INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Revisiting the description of Protein-Protein interfaces. Part II: Experimental study

Frédéric Cazals — Flavien Proust

N° 5501

Février 2005

Thème SYM

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Revisiting the description of Protein-Protein interfaces. Part II: Experimental study

Frédéric Cazals , Flavien Proust

Thème SYM — Systèmes symboliques
Projet Geometrica

Rapport de recherche n° 5501 — Février 2005 — 54 pages

Abstract: This paper provides a detailed experimental study of an interface model developed in the companion article *F. Cazals and F. Proust, Revisiting the description of Protein-Protein interfaces. Part I: algorithms.* Our experimental study is concerned with the usual database of protein-protein complexes, split into five families (Proteases, Immune system, Enzyme Complexes, Signal transduction, Misc.) Our findings, which bear some contradictions with usual statements are the following: (i)Connectivity properties incur important variations across families. These properties are sensitive to water molecules and their variations upon consideration of structural water quantify the filling of packing defects by water molecules. (ii)The model of interfaces as a hydrophilic rim and a hydrophobic core is not general. (iii)At the interface scale, curvature properties correlate with the interface surface area, while locally, the absolute mean curvature follows a bimodal distribution. (iv)About 10% of interfaces consists of several connected components, and this multi-patch structure is independent from structural water. Almost all interfaces feature holes of significant size filled by structural water. Stable crystallographic water molecules play a prominent in reconnecting disconnected interfaces. (v)The chemical composition of interfaces in terms or pairwise contacts features a constant ratio of undetermined interactions, with subtle inter-families variations of determined interactions.

These conclusions shed some light on which structural parameters are most relevant to describe protein-protein interactions, and show that water molecules have a prominent influence on these parameters.

Key-words: Structural Biology, Molecular Modeling, Molecular Interfaces, Voronoï diagrams.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Re-examen des interfaces de complexes Protéine - Protéine. Deuxième partie: Étude Expérimentale

Résumé : Ce travail présente une étude expérimentale précise du modèle d'interface protéine - protéine développé dans le papier *F. Cazals and F. Proust, Revisiting the description of Protein-Protein interfaces. Part I: algorithms.* L'étude porte sur la base de données classique de complexes protéine - protéine, scindée en cinq familles (Protéases, Système Immunitaire,Complexes Enzymatiques, Transduction, Misc.) Les résultats expérimentaux, qui pour certains nuancent des observations classiques, sont les suivants: (i)Les propriétés de connectivité exhibent d'importants changements inter-familles. Elles sont sensibles à l'eau cristallographique, et les variations en présence d'eau quantifient la propension de celle-ci à combler les interstices à l'interface. (ii)Le modèle d'interface comme un *core* hydrophobe d'atomes ayant perdu l'accessibilité au solvant et un *rim* hydrophile n'est pas général. (iii)À l'échelle de l'interface, la courbure moyenne discrète corrèle avec l'aire de l'interface, alors qu'à une échelle locale, la même courbure suit une loi bimodale. (iv)Environ 10% des interfaces ont deux ou trois composantes connexes de taille significative, et cette structure *multi-patch* est indépendante de l'eau cristallographique. Quasiment toutes les interfaces ont des trous de taille variable remplis par des molécules d'eau. (v)La composition chimique des interfaces en termes de paires exhibe une proportion fixe d'interactions non définies, et l'on observe de subtiles variations entre les familles.

Ces observations permettent d'apprécier les paramètres les plus à même de décrire précisément les interfaces de complexes protéine - protéine.

Mots-clés : Biologie Structurale, Modélisation Moléculaire, Interfaces, Protéines, Diagrammes de Voronoï.

1 Introduction

1.1 Understanding the specificity protein-protein interfaces

Non covalent interactions within and in-between proteins are fundamental for the stabilization of protein complexes, and modeling such interactions is mandatory for folding and docking applications. In pursuing this goal, crystal structures of protein complexes provide examples of what nature does, and in spite of several limitations (crystals are static, osmotic conditions of crystallization differ from natural ones, etc) provide a sound basis for learning which interactions are preferred and why. Describing these interactions requires examining the relative positions of atoms, which can be done using a variety of statistics and methods. These statistics are either local and describe properties of one atom and its neighborhood, or global in which case they target global properties of the molecule or complex. Focusing on molecular complexes and their interfaces —rather than proteins in a folding context, we briefly recall here the main statistics and methods, and refer the reader to [CP05] for a comprehensive discussion.

At the atomic level, one finds the identification of interface atoms (i.e. atoms loosing solvent accessibility upon formation of the complex), the classification of these atoms as buried or exposed (often used with a cut-off on the surface area remaining exposed, typically 10^2), the description of neighbors (requires a cut-off distance, usually between 6 and 10 Å), the packing properties using the Voronoi volume (requires a mechanism to bound large or unbounded Voronoi cells —usually explicit water molecules), or the classification of chemical properties of the interface atoms. At a more global level, one encounters the solvent accessible surface area lost upon formation of the complex (SASL), the number of connected components of the interface (based on a clustering of interface atoms), or the characterization of the flatness of interfaces (based on the atomic deviation wrt a least-square plane through the interface atoms).

Motivated by the variety of these statistics and algorithms, we developed a notion of interface unifying these concepts [CP05]. Our construction uses the Voronoi diagram of the atomic balls, and meets four goals: the interface model encodes both local and global properties, provides a way to analyze the geometry and the topology of interfaces, bridges the gap between some standard constructions such as packing properties of interfaces (based upon Voronoi volumes) and the statistics mentioned above, accommodates structural water. As a side effect, our construction also carries two non negligible advantages. First, it provides access to exact quantities rather than approximations. For example, the status (buried / exposed) of an atom reported from the α-complex is certified, while previous calculations based upon approximate values for the surface areas are not. Second, the construction scales as $O(n \log n)$ with n the number of atoms, and calculations of the Voronoi diagram of 10,000 atoms using the CGAL library (www.cgal.org) take about one second on a 2MHz computer.

The goal of this paper is to provide a detailed experimental study of this interface model. To this end, we investigate the interface of 98 complexes selected selected in [CJ99, CMJW03, BCRJ04]. Each complex is analyzed in the following models.

1.2 The AB and ABW models

Low resolution crystals feature crystallographic water molecules. While detailed studies of specific systems usually involve these molecules [BFP95], [Edi00, Chapter 5], global analysis of protein-protein interfaces focused so far on their enumeration and their influence on packing properties [CJ99, CJ02, BCRJ04]. Even though the crystals may have been obtained under conditions of osmotic stress and may not reflect real hydration, it is agreed that these molecules help at improving packing properties at interfaces, and also participate to the formation of hydrogen bond networks.

In developing a notion of interface, one of our goals has been to precisely characterize the role of crystallographic water, which motivates the definition of two models. Before defining them, consider the normalized the B-factor of a water molecule:

$$B_{Water}^{normalized} = \frac{B_{water} - < B_{Group} >}{\sigma(B_{Group})}. \quad (1)$$

The average $<>$ and σ are computed with respect to a group of atoms, which is either the set of all atoms found in the PDB file, or the neighbors of the water molecule of interest. (By neighbor of a water molecule, we refer to all the atoms this molecule is connected to in the Delaunay triangulation.) This second strategy is more local and does not suffer from the influence of loops whose positions fluctuate. Given a threshold (set to one in practice), a water molecule is called stable if its normalized B-factor is less than the threshold. Unstable molecules are discarded, and the remaining ones are simply called stable crystallographic (or structural) water from now on. Depending on the role devoted to structural water, we define:

Model AB Structural water is not incorporated into the Voronoi diagram.

Model ABW Structural water is incorporated into the Voronoi diagram.

Each complex in analyzed in the AB and ABW models. In particular, recall that the AB interface, in the AB or ABW model, consists of the Voronoi facets dual of the AB interface edges. Similarly, the $AW - BW$ interface in the ABW model consists of the facets dual of edges of type AW or BW. Finally, the tricolor ABW interface in the ABW model consists of the union of the two interfaces AB and $AW - BW$.

1.3 Paper overview

Section 2 presents the statistics of interest, while results are discussed in section 3, and put in perspective in section 4. Material and methods are discussed in section 5, while statistics are reported in section 6.

Additional materials are organized as follows. Illustrations of properties highlighted by our interface model are presented in section 7. Tables can be found in section 8, with figures in section 9. The program `intervor` is presented in section 10.

2 Methods: statistics of interest

2.1 Notations

We tag the atoms found in the PDB file with four labels: A and B for the two proteins, W for the structural water, and U for the remaining molecules —co-factors, substrate, metallic ions etc. In all cases, atoms tagged A, B or U are incorporated to the Voronoi diagram. Water molecule are incorporated in the ABW model, but are not in the AB model.

Denote $\#A$ /$\#B$ / $\#W$ /$\#U$ the number of interface atoms of each of the four species. Also denote $\#XY$ the number of interface edges between atoms of types X and Y, and $\#X_{XY}$ the number of interface atoms of type X involved in the XY interface. In the AB model, we always have $\#A_{AB} = \#A$. But in the ABW model, we may have $\#A_{AB} < \#A$ since some A interface atoms may be so due to contacts with interface water molecules.

To compare a statistic S between the two models, we denote $R(S)$ the ratio (S in the ABW model) / (S in the AB model).

2.2 Connectivity

Connectivity of interface atoms is usually measured by the number of neighbors within a distance range. We replace this neighborhood relationship by the notion of *interface neighbors*. Recall that two atoms are interface neighbors if their Van der Waals balls expanded by a water probe intersect, and if this intersection corresponds to an edge in the α-complex of the balls with $\alpha = 0$. Although interface neighbors do not account for all pairs within a distance threshold (i)interface neighbors exactly identify all interface atoms (ii)the most *prominent* pairs are present.

Interface atoms. For the AB model, we report the number of interface atoms of types A and B. For the ABW model, we report the number of interface atoms of types A, B, W, together with the ratios $\#A_{AB}/\#A$ and $\#A_{AW}/\#A$ —and similarly for atoms of type B. These ratios quantify the contributions of the AB and $AW - BW$ interfaces to the qualification of an atom of type A or B as interface atom. For both models, we report the fraction of buried atoms, denoted bur. To compare the two models, we report the ratios $R(\#A + B)$ and $R(bur)$. Finally, we use interface atoms to study the classical model of interface as a core and a rim.

Number of neighbors. In the AB model, we report the average number of neighbors for A and B —given by the formula $n_X = \#XY/\#X$. Combining these values for A and B yields the interface average number of neighbors:

$$n_g = \frac{\#A}{\#A + B} n_A + \frac{\#B}{\#A + B} n_B = \frac{2\#AB}{\#A + B}. \quad (2)$$

In the ABW model, we also report the average number of neighbors for A, that is $n_A = (\#AB + \#AW)/\#A$, and similarly for B. Again, combining these values with the frequencies

of A and B yields the ABW interface average number of neighbors:

$$n_g = \frac{\#AW + \#BW + 2\#AB}{\#A + B}. \tag{3}$$

To study the potential asymmetric situations, we also report $r_{Mm} = Max(n_A, n_B)/min(n_A, n_B)$. Finally, to compare the models AB and ABW, we report the ratio $R(n_g)$.

2.3 Topology and geometry

The only reports we are aware of on the topology of interfaces are [CJ02] and [BER04]. Based on interface atoms, the average linkage clustering algorithm [CJ02] reports connected components. This clustering algorithm favors *globular* clusters but does not provide additional information on the geometry of the clusters and on solvent accessibility. On the other hand, the interface of [BER04] does exploit geometric and topological properties of the diagram of balls, but the atoms involved are not the atoms loosing accessibility in the complex. Our construction enjoys both properties and allows a precise study of the geometry and of the topology of interfaces.

Voronoi Interface Surface Area. Given a Voronoi interface, the Voronoi Interface Surface Area (VISA) is naturally defined as the sum of the surface areas of the facets defining the interface. Denote $Area(XY)$ the surface area of the bicolor interface XY. To compare VISA with Solvent Accessible Surface Area Lost (SASAL), we define the following quantities:

- $\mathrm{A}_{\mathrm{Ref.}}$. The SASL from [CJ99, CMJW03]. The SASL is computed without structural water, and is not exact since computed by sampling the atomic balls. But it provides a good approximation of the exact SASL and is used as reference.
- $\mathrm{A}_{\mathrm{BER}}$. The VISA reported in [BER04].
- A_{AB}; AB model. Defined as the sum of the surface areas of the AB facets, that is $\mathrm{A}_{AB} = Area(AB)$.
- A_{ABW}; ABW model. In the ABW model, we have AB but also AW and BW facets. Since a water molecule is always *sandwiched* between atoms of type A and B, it is natural to define the VISA by $Area(ABW) = Area(AB) + (Area(AW) + Area(BW))/2$.

To perform comparisons, we report the values of the surface areas just defined, together with the coefficients of the linear regression $\mathrm{A}_{\mathrm{BER}}$ vs $\mathrm{A}_{\mathrm{Ref.}}$, A_{AB} vs $\mathrm{A}_{\mathrm{Ref.}}$, A_{ABW} vs $\mathrm{A}_{\mathrm{Ref.}}$.

Connected components. Recall that an edge connected component (cc for short) of the AB interface consists of a collection of AB facets sharing a Voronoi edge. Given a threshold $t_s \in (0, 1)$, define a significant connected component (scc) as a cc whose surface area is at

least a fraction t_s of the VISA of the AB interface. In the AB model, we report the number of cc and scc, denoted $\#cc$ and $\#scc$.

Consider now the ABW model. As explained in [CP05], we merge the edge connected components of the AB and $AW - BW$ interfaces. We therefore report the number of cc and scc before and after the merge process, denoted $\#cc_{bm}$, $\#scc_{bm}$, $\#cc_{am}$, $\#scc_{am}$. By definition, the number of cc and scc in the ABW model before the merge process corresponds to AB connected components only —i.e. components of the $AW-BW$ interface are ignored.

Flatness and curvature properties. To report on curvature properties, we focus on the absolute discrete mean curvature of the AB interface. The incentive for using the AB interface is that Voronoi interface edges are manifold —which may not the case for the ABW interface. Notice that we measure the curvature of the bicolor interface, rather than computing a curvature notion associated to the atoms themselves. However, it should be kept in mind that each Voronoi facet witnesses the intersection of two atomic balls whose centers are located apart from the plane containing the Voronoi facet, so that the curvature of the interface is a fingerprint of the *curvature* of the union of interface atoms.

Denote $s_H = \sum_{e \text{interior Voronoi edge}} \mid h(e) \mid$ the total absolute mean curvature, with $h(e) = \beta(e)l(e)$. (Recall that $l(e)$ is the length of edge e and $\beta(e)$ the dihedral angle at edge e.) To get a global view of the interface mean curvature, we focus on s_H and its relationship to the VISA A_{AB}. To get a local view of curvature properties, we report the expectation and the standard deviation of the absolute value of the dihedral angle —the weight of the angle at edge e being $l(e)/L$, with L the sum of the lengths of all internal edges.

2.4 Chemical composition of interfaces

The chemical characterization of interfaces is usually carried out by cumulating the SAS lost by the interface atoms as a function of the chemical type of the atoms. While this analysis gives a global view of the interface, it does not account for pairwise interactions. To get such a characterization, we annotate the interface edges, and charge the surface areas of the corresponding interface facets to the annotated type —see section 5.3. Notice that in doing so, one encounters undetermined facets —a problem intrinsic to any method focusing on pairwise interactions when the two parties are not *compatible*.

3 Results

In this section, we detail the results for the three groups of statistics —see tables in section 6. The reader is also invited to visit `www-sop.inria.fr/geometrica/team/Frederic.Cazals/intervor/index.html` for a visual inspection of the interfaces.

3.1 Interface atoms and connectivity properties

Number of interfaces atoms. Figure(s): 10. The number of interface atoms in protein-protein complexes has been investigated in [RJ98b, CJ02], where it is shown that this number correlates with the interface surface area SASL. To *calibrate* our study, we comment briefly on the number of interface atoms found in our study.

We first compare the standard values [CJ02] in the AB model. The two values are always within 20 percents. Our calculations involve of the atoms found in the PDB file, that is we do not exclude atoms having a null temperature factor or an occupancy factor less than one. The incentive for doing so is that these atoms are present in the molecule, and removing them would create holes at the interface. In any case, we found that the number of such atoms at the interface is always less than 5% excepted for 1dan 8.75% and 1mkw 22.62%. Therefore, the discrepancies observed come from different sources which are the following. First, our definition of interface atom is strict. An atom losing SAS area in the complex is declared at the interface —which differs from [CJ02] where the loss must be more than 0.1^2. Second, while the solvent accessibility of the previous studies are reported using an approximate algorithm [Hub92], we use the exact information encoded in the α-complex of the atomic balls.

Atoms, neighbors and buried atoms. Figure(s): 11, 12. We discuss at once the number of interface atoms, the number of buried atoms, and the connectivity properties in both models. Comparing the AB and ABW models, one observes that:

- $\#A+B$: the number of interface atoms gets multiplied on average by 1.21, the five groups having comparable statistics —mean values of $R(\#A+B)$ between $[1.18, 1.26]$.
- n_g: the average number of neighbors, which is rather heterogeneous in the AB model —proteases have an average of 3.56 while other averages are in the range $[3.36, 3.46]$, gets more uniform in ABW —all averages in the range $[3.24, 3.36]$.
- bur: the ratio of buried atoms is rather heterogeneous in both models, and incurs a shift of about 0.1 from AB to ABW.

To understand these observations, consider interface atoms in the AB model, and assume one or several water molecules squeeze in-between them. These water molecules *brake* interface edges of type AB, which get replaced by edges of types AW and BW. The atoms of A and B loose accessibility and might get buried. The average number of neighbors of these atoms decreases. Indeed, water molecules are closer from the atom whence fewer neighbors —-the surface where their centers lies has smaller surface area than that of the original neighbors. Thus, the variation of n_g and of the ratio bur provides a measure of the packing defects filled by the water molecules. Consider now interface atoms in the ABW model which are not so in the AB model. These are atoms recruited by the water molecules either at the interface periphery or in regions of loose packing.

Remarkable complexes for $R(\#A+B)$ are 1vfb (2.28), 1osp (1.62), 1l0y (1.58), 1tco (1.69). Interestingly, for the first three interfaces, water molecules are filling are filling *creeks* at the interface. The situation is radically different for 1tco where the water molecules reconnect 5 significant connected components —see section 3.4. Remarkable complexes for $R(bur)$ are 1dan (2.17), 1kxv (2.46), 1tco (2.22). For example, the interface of 1kxv features an inner ring filled by 11 water molecules.

To get a more detailed view of the number of neighbors, consider the ratio r_{Mm}. Proteases and Signal transduction have the largest and smallest values, respectively 1.45 and 1.06 in the AB model. (All values of this statistic incur a negligible shift of about 0.01 in the ABW model.) As illustrated by Fig. 6, the specificity of most of proteases comes from the recognition by the enzyme in a deep pocket of an Arginine side chain of the substrate / inhibitor.

Additional observations are the following. For Proteases, the ratio r_{Mm} correlates negatively to the interface size [1]. The statistic n_g does not correlate with the interface size in any of the groups. Proteases and Signal transduction form the two most homogeneous groups. In any case, the values of n_g reported are comparable to the most frequent numbers of hydrophobic and hydrophilic contacts (for an atom either hydrophobic or hydrophilic) reported in [RJ98a]. Incidentally, an exceptional complex illustration of the high value of n_g is provided by the tax viral peptide in complex 1a07 —Fig 3, which is completely buried in-between HLA-A2 and a TCR.

On the core and the rim. The usual interface model consists of a rim of atoms retaining solvent accessibility surrounded a core of buried atoms [CJ99]. Buried atoms represent in average a fraction of .33 (0.42) of the total number of interface atoms in the AB (ABW) model. In both models, proteases are remarkable since their average (0.4) is 0.1 above those of the other classes. Even when water molecules are included, the ratio of buried vs exposed never exceeds two. If a core (modeled as a surface patch) were bounded by a rim (modeled as a necklace), the number of atoms of the former would vary as the square of the later.

This suggests that the model of interfaces as core-rim is more contrasted, which means that one finds exposed atoms in the *middle* of the interface, with possibly buried atoms at the interface *boundary*. (These atoms may be buried due to contacts with atoms of the same protein.) This is illustrated on Figs. 3 and 4. Visual inspection of complexes actually shows that for those complexes which do not exhibit a *clear* core and rim, one usually finds *clusters* of buried atoms at one or several locations of the interface. In other words, the interfaces are usually fragmented in terms of exposed/buried status of the interface atoms. This observation is not very surprising given the existence of hot spots which usually feature a few buried amino-acids [BT98]. Providing a quantitative measure of this phenomenon is possible by studying for each atom the topological patterns made by its neighboring atoms (i.e. the atoms of the same molecule).

[1] This property is likely to be related to the fact that anchor residues bear some independence with the set of all interface residues, whose size may vary.

3.2 Area values

Reference SASL. Figure(s): 13. Recall that our interface calculations correspond to a probe of radius $r_w = 1.4$, together with a filtering of interface edges with $M = 25$. Also recall that the areas reported by Janin et al. [CJ99] serve as reference for SASL and span the range $[565, 2330]$ Å^2 with an average of 952, while values reported by Ban et al. [BER04] span the range $[397, 2408]$ with an average of 963 [2].

SASL versus VISAs. Figure(s): 14(a-c). In the AB model, we find that $\mathrm{A}_{AB} \in [724, 2959]$ with an average of 1261, while in the ABW model, $\mathrm{A}_{ABW} \in [779, 3634]$ with an average of 1445. To compare the different models, we first look at the correlations between the VISAs and the reference SASL. We first analyze the AB model. As reported in [BER04], the linear correlation coefficient between $\mathrm{A}_{\mathrm{BER}}$ and $\mathrm{A}_{\mathrm{Ref.}}$ is of 0.86 and corresponds to a slope of 0.74. (Since 70 out of 98 VISA values are reported [BER04], the correlation coefficient corresponds to this subset of interfaces.) The linear correlation between VISA A_{AB} and SASL values ($\mathrm{A}_{\mathrm{Ref.}}$) is equal to 0.89 —with a slope of 1.57. These results show that the AB interface is better at capturing information on SASL, and that the surface areas of the interface Voronoi facets provide a sound geometric quantity to encode the neighborhood relationship between interface neighbors.

Addition of $AW-BW$ interface facets makes the linear correlation between VISA A_{ABW} against $\mathrm{A}_{\mathrm{Ref.}}$ shift to 0.88 —slope 1.76, an observation expected since the calculation of VISA ABW includes water molecules while SASL does not. To further quantify the role of water molecules in terms of surface area, we also computed the ratio between surface areas of the AB facets in the ABW model wrt to Area(ABW). These ratios span the range $[1, 2.45]$, and corresponds to a linear correlation coefficient of 0.84 with a slope of 0.67, assessing the major contribution of water molecules in direct contacts between atoms.

3.3 Curvature properties

It is usually admitted that interfaces tend to be flat [JT96, CJ02], a statement based upon the average deviation (2.5Å) of interface atoms to the least square plane through their centers.

Global curvature. Figure(s): 15(a). A first observation about the global curvature statistic s_H is that the overall picture contrasts with the previous statement. There is an obvious correlation between the interface surface area and s_H, with a linear regression coefficient of 0.96. This correlation is the main property and is valid within the five families —the standard deviation of $s_H/Area(AB)$ lie in the range $[0.07, 0.11]$. Since our curvature measure involves Voronoi edge lengths and dihedral angles, assuming some uniformity on the dihedral angles, the correlation between the surface area and s_H expresses the fact that the sizes of the Voronoi facets and their boundary length are rather uniform. Moreover, since facets bend across Voronoi edges, larger interface have the ability to develop bent shapes —although the amount of absolute mean curvature by element of surface area remains constant. As an

illustration of curvature properties, the reader may consult complex 1ppe on Fig. 5, which features an interface with a deep pocket, and complex 2trc on Fig. 6 —interface with a *right* angle.

Local curvature properties. Figure(s): 15(b), 16(a-f). To get more insight on local curvature properties, we focus on the absolute value of the dihedral angle $\beta(e)$ and on the edge lengths $l(e)$. Its average and standard deviation are remarkably uniform. To analyze this parameter further, we plotted the distribution of the angle and compared them to those of the length of interface edges. We picked three complexes belonging to the three categories flat ($s_H \leq 25$) interface , moderately curved ($s_H \in [25, 45]$, and curved ($s_H \geq 45$). The complexes are 1a0o, 1udi 2trc and exhibit identical behaviours. While the distribution of edge lengths is unimodal, that of the angle resembles a bimodal distribution with a peak and a slowly uniformly decreasing plateau. Evaluation of a discrete entropy ($H = \sum_i p_i \log p_i$) for both statistics shows that $H(l(e)) \sim 2.1$ and $H(\beta(e)) \sim 2.25$. Recall that a dihedral angle between two Voronoi facets involves three atoms. The distance between two atoms and the dihedral angle respectively have one and two degrees of freedom, so that the observation on entropy is expected. This also shows that the informations encoded are different, which is of interest for geometric statistical potentials.

3.4 Connected components

Figure(s): 17, 18.

In dealing with connected components, one faces two types of informations: the number of cc and their relative size. To get rid of small regions of the interface, we focus on significant cc as defined in section 2.3. Practically, we use $t_s = 0.1$. As reported in Table 4, the number of cc in both models is small and varies between one and nine. For example, it is found that in the AB model, $\#cc \in 1..6$, while $\#scc \in 1..3$. To fully understand these values, notice that $\#cc_{bm} \geq \#cc$ since water either create new components or split components already present. One can also expect $\#cc \geq \#cc_{am}$ —if water molecules bridge more gaps than they create new components. In general, these relationships do not translate onto their counterparts on significant cc. For example, one cannot expect $\#scc_{bm} \geq \#scc$ since water molecules may have split significant cc into non significant cc. In spite of these observations, we shall see that all complexes featuring multiple patches satisfy

$$\#scc_{bm} \geq \#scc \geq \#scc_{am}, \tag{4}$$

and that the types $(=, >)$ of the two binary operators qualify the scale at which water molecules operate.

Complexes with a multi-patch structure. Using $t_s = 0.1$, the number of complexes with significant cc are respectively 16 in the AB model, 19 in the ABW model before the merge, and 8 after the merge. In the ABW model, addition of water therefore causes 11 complexes to loose a multi-patch structure. The eight complexes featuring between two and

three significant components after the merge process are —in parenthesis the number of components and the number of interface water molecules: proteases { 1mkw (2, 9), 1tbq (2, 7)}, Immune system { $\emptyset$ }, Enzyme Complexes {1djf (3, 7)}, Signal transduction {1got (2, 27), 1aip (2, 2), 1efu (2, 65) }, Misc {1fq1 (1, 0), 1dkg (3, 4)}. For all these complexes but 1efu, the multi-patch structure is actually independent from the water molecules —the three numbers of significant cc agree. (For 1efu, we have $\#scc_{bm} = 4 > \#scc = 3 > \#scc_{am} = 2$.) This observation essentially shows that multi-patch structures are well separated and are not merged upon consideration of structural water.

Water bridging gaps and packing defects. Having discussed the multi-patched complexes, consider the 11 cases where the multi-patch structure disappears upon consideration of water molecules. As just discussed, all multi-patch complexes but 1efu are characterized by $\#scc_{bm} = \#scc = \#scc_{am}$. The remaining 11 complexes correspond to the three remaining cases when replacing the $\geq$ operator by either $>$ or $=$ in Eq. (4):

- $\#scc_{bm} > \#scc = \#scc_{am}$. The number of cc in the AB equals that in the ABW model after the merge. In other words, water molecules fill small gaps which are already connected in the AB model. This case represents $2/11$ complexes (1cse, 1tx4).
- $\#scc_{bm} = \#scc > \#scc_{am}$. The number of scc in the AB equals that in the ABW model before the merge. In other words, there are significant gaps which are not filled if water molecules are not there. This case represents $8/11$ complexes (1dan, 1wej, 1osp, 2pcc, 1gg2, 1l0y, 1ycs, 1hwg).
- $\#scc_{bm} > \#scc > \#scc_{am}$. The number of scc in the AB is in-between the numbers before and after the merge process. This situation is a mix of the to previous ones, and corresponds to a single complex (1tco). (And also to 1efu in complexes with a multi-patch structure.)

For these complexes, water molecules either partially or totally reconnect disconnected interfaces. A typical example is provided by 1gg2 which features two patches around a *hinge* disconnected by a *tunnel*. This tunnel is filled by 11 water molecules (out of the 29 structural water molecules) which all retain accessibility to the solvent. The merge phenomenon is more radical for three complexes which 1dan, 1efu and 1tco. For that later complex, water molecules connect five components of the AB interface into a single one. Incidentally, 1tco is also the complex where the consideration of water molecules yields the most significant increase in the number of interface atoms.

Connected components versus loops and nets. Having discussed connectivity, let us focus on a connected component of the AB interface. The component is either be a topological disk or is non simply connected if it features one or several hole(s). To quantify this property, we computed in the ABW model, for each significant cc of type AB, the number of significant boundary loops before the merge process and the number of significant

nets [2] in the ABW model after the merge process. (By significant loop or net, we refer to whose length (exposed length for nets) is at least some fraction t_s of the overall boundary (net) length of the interface.)

Setting $t_s = 0.1$ —as for surface area, 14 complexes feature one scc with at least one significant hole —in parenthesis the number of holes for a scc having more than one hole together with the number of water molecules in each hole: Protease: 1cho, 1acb, 4htc; Immune system: 1bvk 1kxt, 1kxv, 1mel, 1nfd, 1kb5 (3; 1, 2, 3); Enzyme Complexes: 1udi (3; 1, 1, 5), 1ugh (3; 1, 1, 7); Signal transduction: 1a2k (3; 1, 1, 1), 1got, 2trc (2; 1, 6). As an illustration, 1kxv contains a particularly impressive hole featuring 11 water molecules.

Comparison to previous analysis. As already pointed out, a study of the number of patches has been carried out in [CJ02] using an average linkage clustering algorithm. The results are in pretty good agreements for interfaces with 3 or 4 patches —apart from 1kb5 the number of patches agree with $\#scc$. But for two patch interfaces, we found the average linkage clustering algorithm overestimates the number of patches. (This behavior is actually expected since the average linkage clustering algorithm, based on average distances, favors globular clusters.) As an illustration, the reader may consult 1iai or 1kb5 which feature compact arrangement of interface atoms.

3.5 Chemical composition

So far, the chemical characterization of interfaces has focused on atoms rather than contacts. In particular, [CJ99] noticed that the contribution to the SASL from non-polar, polar and charged groups are respectively of 56%, 29%, 15%.

In terms on interactions, recall that we use eight types to annotate interface edges of the ABW interface —Table 2.

We first observe that examining the composition of interactions in the AB model is misleading since hydrophobic interactions dominate polar ones —0.35 vs 0.2. Addition of structural water balances these coefficients and hydrophobic versus polar contacts become comparable —0.28 vs 0.26. The contribution of polar interactions with water is homogeneous, and is about 0.1 of the total surface area, which is about 1/6 of the determined interactions. Speaking of which, the ratio of undetermined contacts is about 0.4 in both models. It is appealing to believe that well-defined interactions are sufficient to describe the interface, despite the high proportion of undetermined interactions. Confirming this hypothesis would require going through hot spots and checking that the ratio of undetermined contacts in these areas is less than the average. Apart from these observations, few prominent properties emerge. Proteases (Signal transduction) are characterized by a low (high) ratio of charged groups. As observed in [CJ99], these observations are likely to be related to the size and the solubility of the components of the complexes. Complexes from the immune system have the highest ratio of polar interactions.

[2] Recall that a network is the boundary of a cc in the interface ABW. In general, a network is not a one-manifold.

Finally, it should be observed that the pairwise chemical composition is linked to the geometry and the topology of the interface. One finds large and continuous hydrophobic patches, generally surrounded by polar interactions. As an illustration, the reader may examine the deep creeks of 1vfb —filled by water molecules concentrating polar interactions.

4 Conclusion

4.1 Conclusion

To conclude, we briefly recap the main features of the complexes studied. Proteases are characterized by the largest ratio of buried atoms and by the fact that the two proteins play asymmetric roles in terms of number of neighbors. Complexes from the immune system have the least and most homogeneous interface size —in terms of number of atoms. Moreover, this group is the only one where all the interfaces have a single significant connected component, although water molecule often fill creeks about the loops in contact. In terms of number of atoms, signal transduction complexes have the largest and and least homogeneous interfaces. This group also features the largest number of multi-patch interfaces. Finally, these complexes also characterized by a low average (and standard deviation) number of neighbors, as well as a rather symmetric role played by the two proteins.

May be the most salient feature of our interface model is to highlight the role of structural water. Water molecules increase the ratio of buried atoms (thus filling packing defects) and decrease the average number of neighbors (atoms in direct contact.) Analysis of connected components shows that 19/98 complexes have at least two significant patches if water is excluded, a number dropping to 8/98 if water is considered. While multi-patch interfaces are independent from structural water (apart from one complex), water molecules reconnect 11/98 interfaces having connected components separated by small gaps. In 14 cases, the topology of the interface is non trivial since the main connected component features between one and three holes, each filled by a number of water molecules between one and eleven. Another interesting observation is that the curvature of interfaces does not exhibit any specific property within the groups. Global curvature properties correlate to the interface surface area, while local properties seem to follow well-defined distributions —dictated by the chemistry.

As a conclusion, we believe this study validates an interface model which is, to the best of our knowledge, the only one to provide a chemical, geometric and topological description of interfaces, from the atomic scale to the interface scale.

4.2 Future work

In this paper, we focused on statistics at the interface scale so as to validate our interface model. The natural follow-up is to examine statistics at the residue and atomic levels. Since out interface has the unique advantage to encode precise local properties of the arrangements of atoms, the following topics should be investigated.

Core, rim, interactions, hot spots. We have shown that the model of interfaces as a core and a rim is not general, and we have also shown that the composition of pairwise interactions incurs subtle variations. Properties of interfaces at the atomic and residue scales (rotamers) should be quantified from a geometric and topological standpoint —cf also our definition of interface core. This analysis should expand the understanding of hot spots, and provide specificities of protein-protein complexes families.

Packing properties. We showed how connectivity informations encode packing defects. The connexion between our interface and packing properties has to be done. Such a contribution should bridge the gap between the Voronoi volumes of Lee-Richards, the surface complementarity measure of Lawrence and Colman, the gap index of Laskowski, and the the pockets of Edelsbrunner.

Flexibility issues. We focused on a static analysis of interfaces. Studying the same properties in a dynamic context is a must, and should provide insights on flexibility issues in complex formation.

Scoring functions. The geometric and topological properties of the α-complex pave the way to develop parameterizable models accommodating a notion of geometric entropy, which might be related to (approximate) free energies of the systems studied. These models should integrate local properties on the interface —amino-acids or atoms involved in a region of the interface.

References

[BCRJ04] R.P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Bio.*, 336, 2004.

[BER04] Y.-E. A. Ban, , H. Edelsbrunner, and J. Rudolph. Interface surfaces for protein-protein complexes. In *RECOMB*, 2004.

[BFP95] B.C. Braden, B.A. Fields, and R.J. Poljak. Conservation of water molecules in an antibody-antigen interaction. *J. of Molecular Recognition*, 8, 1995.

[BT98] A.A. Bogan and K.S. Thorn. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280, 1998.

[CJ99] L. Lo Conte and J. Janin. The atomic structure of protein-protein recognition sites. *J. Mol. Bio.*, 185, 1999.

[CJ02] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47, 2002.

[CMJW03] R. Chen, J. Mintseris, J. Janin, and Zhiping Weng. A protein-protein docking benchmark. *Proteins*, 52:88–91, 2003.

[CP05] F. Cazals and F. Proust. Revisiting the description of protein-protein interfaces. part i: algorithms. Technical Report 5346, INRIA, 2005. Revised version.

[Edi00] C. Kleanthous Editor. *Protein-protein recognition.* Oxford, 2000.

[Hub92] S. Hubbard. Access: a program for calculating accessibilities. Technical report, Univ. college of London, 1992. http://wolf.bms.umist.ac.uk/naccess.

[JT96] S. Jones and J. Thornton. Principles of protein-protein interactions. *PNAS*, 93(1), 1996.

[RJ98a] C. Robert and J. Janin. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J. Mol. Bio.*, 283, 1998.

[RJ98b] C.H. Robert and J. Janin. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J. Mol. Biol.*, 283(5), 1998.

5 Appendix: methods

5.1 Data set of protein-protein complexes

Our study involves the complexes investigated in [CJ99, CMJW03, BCRJ04]. According to the header of the original PDB file, the atoms of each complex are tagged with four labels: the two proteins A and B, the water molecules determined by X-ray crystallography (W) and the remaining molecules (cofactors, substrate, ...) (U). To avoid redundancy, we selected a single representation of each protein in the crystallographic unit.

Interface edges are filtered with $M = 25$. This value is sufficient to get rid of large Voronoi facets for three complexes —-1ydr, 1tbq and 1cgi. Discarding these facets is a convenient and inexpensive way to make the VISA AB and SASL coherent. Notice though, that the right way to encode a surface area information on the facets consists of using the weights defined in [CP05].

5.2 Stable crystallographic interface water molecules

Our filtering strategy for water molecules is twofold. First, water molecules are filtered by their normalized B-factor —Eq. (1), and a threshold of one has been used in practice. Second, since we investigate interfaces, we discard all the molecules that do not make at least one contact with each protein —see the definition of interface water in [CP05].

5.3 Annotations of atoms and atoms pairs

Interface atoms: buried versus exposed. In a complex featuring two molecular species A and B, an interface atom is an atom loosing accessibility to the solvent upon formation of the complex. Moreover, the atom is called buried if it completely looses accessibility, and exposed otherwise. We use this classification in our AB model. The situation gets more clumsy in the ABW model. Since we include structural water, the atoms of A (B) may loose exposure to the solvent (i.e. free water molecules) due to atoms of B (A), but also due to structural water. An exposed interface atom in the AB model can therefore get buried in the ABW model, while being connected to structural water reducing its accessibility.

Chemical profile of atoms of interface edges. Each heavy atom present in the PDB file is represented by a Van der Waals ball. As the PDB atom notation depends on the atom position in the residue and not on its chemical profile, we use the atom types described in Table 1 instead. This atom notation may not be as fine as those provided by SATIS or SYBYL, but it is sufficient to account for potential polar interactions between peptidic bonds of hydrophobic amino acids or non-polar contacts between aliphatic carbons of lysine side-chains for example. Notice also that more accurate classifications would increase the combinatorics of interactions, making results difficult to interpret. For non-peptidic residues, we used the first letter of the atom name as element type. All these atoms are inserted in

the Voronoi diagram. But since we focus on interfaces, the only such atoms of interest are the oxygens of water molecules.

To annotate two-body interactions corresponding to interface edges, we use the profiles of the vertices as specified in Tables 2 and 3. This classification can naturally be coarsened as hydrophobic Hy (types C, Ca), polar Po (P, Pa, Pw), charged Ch (P+), and undetermined Un (U, Uw) interactions so as to match previous classifications.

Atom class	Symbol	Color	Atoms in proteins	Other Atoms
Aliphatic carbon atoms	C	yellow	C_α, aliphatic carbon ($C_{\beta,\gamma,\delta,\ldots}$)	generic carbon
Aromatic carbon atoms	Ca	orange	aromatic carbon in phenyl, phenol, imidazol and indol groups	
Nucleophilic atoms (electron donor)	P	green	oxygen in serine, threonine and phenol groups, sulfur in thiol groups	generic nitrogen, oxygen and sulfur
	Pw	green		oxygen in water
Unsatured nucleophilic atoms	Pa	dark green	nitrogen in amide, imidazol and indol groups	
Charged atoms	P+	red	nitrogen in guanidium group (N_ϵ) or in amine group	
Others	U	gray		phosphorus atoms, sulfur oxides, halogen atoms, metal atoms, and others

Table 1: Chemical classification of atoms used to annotate vertices in the Delaunay diagram and the α-complex.

Type	Symbol	Color
Aliphatic	C	yellow
Aromatic	Ca	orange
Polar	P	light green
Unsaturated	Pa	green
Charged (salt-bridge)	P+	red
Undefined interactions	U	gray
Water Polar	Pw	olive green
Water Contact	Uw	light gray

Table 2: Annotation of interface edges: the height types

	C	Ca	P	Pa	P+	U	Pw
C	C	C	U	U	U	U	Uw
Ca	C	Ca	U	Ca	U	U	Uw
P	U	U	P	P	P	U	Pw
Pa	U	Ca	P	Pa	P	U	Pw
P+	U	U	P	P	P+	U	Pw
U	U	U	U	U	U	U	Uw
Pw	Uw	Uw	Pw	Pw	Pw	Uw	Pw

Table 3: Annotation of interface edges

6 Main statistics

In this section, we provide an overview of the main statistics, together with statistics for the five groups. For convenience, we use the following abrevations: Proteases (P), Immune system (IS), Enzyme complexes (EC), Signal transduction (ST), Misc (M).

6.1 Overview

Model	Variable	min	max	mean	statd	median
AB	$\#(A+B)$	118	562	232.40	87.82	206
ABW	$\#(A+B)$	138	745	282.63	113.01	255.5
Cmp	$R(\#(A+B))$	1.	1.69	1.21	0.18	1.21
AB	n_g	2.91	4.08	3.46	0.19	3.50
ABW	n_g	2.58	3.99	3.32	0.22	3.33
Cmp	$\mathrm{R}(n_g)$	0.86	1.03	0.95	0.03	0.95
AB	r_{Mm}	1.	2.77	1.23	0.26	1.14
ABW	r_{Mm}	1.	2.9	1.23	0.27	1.14
Cmp	$\mathrm{R}(r_{Mm})$	0.81	1.16	1.00	0.06	1.
AB	bur	0.14	0.51	0.33	0.08	0.33
ABW	bur	0.17	0.68	0.42	0.11	0.43
Cmp	$R(bur)$	0.97	2.4	1.29	0.31	1.20
AB	A_{AB}	724.62	2959.13	1261.75	488.88	1093.45
ABW	A_{ABW}	779.59	3634.20	1445.42	574.22	1285.17
AB	$s_H/Area(AB)$	0.75	1.42	1.02	0.10	1.01
AB	$\#cc$	1	6	1.98	1.20	2
AB	$\#scc$	1	3	1.20	0.49	1
ABW	$\#cc_{bm}$	1	9	2.20	1.47	2
ABW	$\#scc_{bm}$	1	5	1.27	0.67	1
ABW	$\#cc_{am}$	1	5	1.56	0.93	1
ABW	$\#scc_{am}$	1	3	1.10	0.36	1

Table 4: Overview of the main statistics

6.2 Pairwise chemical composition of interfaces

Refer to section 5.3 for the specification of the interaction types.

	All	P	IS	EC	ST	M
Hy	0.35	0.37	0.32	0.35	0.35	0.36
Po	0.20	0.20	0.25	0.20	0.16	0.17
Ch	0.04	0.02	0.04	0.05	0.08	0.06
Un	0.39	0.39	0.37	0.39	0.40	0.39
C	0.27	0.30	0.22	0.28	0.28	0.28
Ca	0.08	0.06	0.10	0.07	0.07	0.08
P	0.15	0.14	0.20	0.15	0.12	0.12
Pa	0.04	0.06	0.04	0.04	0.03	0.04
P+	0.04	0.02	0.04	0.05	0.08	0.06
U	0.39	0.39	0.37	0.39	0.40	0.39
Pw	0.	0.	0.	0.	0.	0.
Uw	0.	0.	0.	0.	0.	0.

Table 5: Chemical composition of interfaces: model *AB*

	All	P	IS	EC	ST	M
Hy	0.28	0.29	0.27	0.28	0.27	0.31
Po	0.26	0.27	0.29	0.25	0.23	0.22
Ch	0.03	0.01	0.02	0.04	0.06	0.05
U	0.40	0.41	0.39	0.41	0.42	0.40
C	0.22	0.23	0.18	0.23	0.21	0.24
Ca	0.06	0.05	0.08	0.05	0.05	0.06
P	0.11	0.10	0.15	0.11	0.08	0.09
Pa	0.03	0.04	0.03	0.03	0.01	0.03
P+	0.03	0.01	0.02	0.04	0.06	0.05
U	0.30	0.28	0.30	0.30	0.29	0.31
Pw	0.11	0.13	0.10	0.11	0.12	0.09
Uw	0.10	0.13	0.09	0.11	0.13	0.08

Table 6: Chemical composition of interfaces: model *ABW*

6.3 Statistics by groups

We report the following statistics (mean, max, mean, standard deviation (stdd), median (med)) by complex family. The five lines respectively correspond to the five groups. For variable $S \in \{\#A+B, bur, r_{Mm}, n_g\}$, the third table presents the statistics of $R(S)$.

6.4 Number of interface atoms #A+B

min	max	mean	stdd	med
144	430	230.75	83.56	204.
163	287	206.89	35.08	203.
118	376	235.09	70.68	233.
138	562	317.58	140.54	296.5
127	527	217.05	91.05	205.

AB

min	max	mean	stdd	med
144	588	287.67	99.75	259.
163	406	246.28	57.64	237.
151	473	280.72	85.29	272.
138	745	400.83	180.64	402.
145	636	255.21	112.84	221.

ABW

min	max	mean	stdd	med
1.	1.54	1.25	0.14	1.27
1.	1.62	1.19	0.20	1.14
1.	1.48	1.20	0.15	1.22
1.	1.57	1.26	0.18	1.27
1.	1.69	1.18	0.20	1.13

Comparison: ABW / AB

6.5 Ratio of buried atoms *bur*

min	max	mean	stdd	med
0.17	0.51	0.40	0.08	0.42
0.17	0.51	0.31	0.07	0.31
0.14	0.47	0.29	0.08	0.3
0.24	0.36	0.31	0.03	0.33
0.18	0.42	0.31	0.06	0.32

AB

min	max	mean	stdd	med
0.26	0.63	0.49	0.08	0.49
0.17	0.6	0.41	0.12	0.40
0.21	0.6	0.39	0.12	0.39
0.24	0.68	0.42	0.11	0.44
0.22	0.52	0.36	0.07	0.38

ABW

min	max	mean	stdd	med
1.	2.17	1.25	0.24	1.18
0.99	2.46	1.33	0.39	1.27
0.98	1.92	1.37	0.34	1.29
1.	1.93	1.31	0.28	1.28
1.	2.22	1.22	0.32	1.08

Comparison: ABW / AB

6.6 Average number of neighbors n_g

min	max	mean	stdd	med
3.13	3.84	3.56	0.13	3.57
3.13	4.08	3.46	0.20	3.46
3.02	3.65	3.38	0.22	3.52
3.12	3.59	3.36	0.14	3.39
2.91	3.72	3.41	0.23	3.48

AB

min	max	mean	stdd	med
2.89	3.59	3.32	0.15	3.33
3.11	3.99	3.36	0.22	3.34
2.9	3.61	3.29	0.22	3.25
3.	3.59	3.24	0.17	3.23
2.58	3.72	3.30	0.32	3.37

ABW

min	max	mean	stdd	med
0.86	1.	0.93	0.03	0.92
0.89	1.03	0.97	0.03	0.97
0.91	1.00	0.97	0.03	0.97
0.90	1.	0.96	0.03	0.97
0.88	1.	0.96	0.03	0.96

Comparison: ABW / AB

6.7 Asymetricity of the number of neighbors r_{Mm}

min	max	mean	stdd	med
1.14	2.07	1.45	0.23	1.4
1.01	2.77	1.19	0.31	1.13
1.	1.39	1.14	0.12	1.13
1.	1.14	1.06	0.04	1.06
1.	1.49	1.12	0.13	1.08

AB

min	max	mean	stdd	med
1.01	2.13	1.44	0.25	1.39
1.02	2.9	1.18	0.34	1.12
1.02	1.49	1.18	0.14	1.14
1.02	1.19	1.08	0.04	1.08
1.	1.54	1.13	0.15	1.08

ABW

min	max	mean	stdd	med
0.81	1.16	0.99	0.07	1.
0.86	1.06	0.98	0.04	1.
0.90	1.14	1.02	0.06	1.02
0.92	1.14	1.01	0.05	1.00
0.89	1.09	1.01	0.05	1.

Comparison: ABW / AB

6.8 Number of significant connected components

min	max	mean	stdd	med
1	3	1.14	0.44	1.
1	2	1.07	0.26	1.
1	3	1.27	0.64	1.
1	3	1.41	0.66	1.
1	3	1.31	0.58	1.

AB

min	max	mean	stdd	med
1	3	1.17	0.47	1.
1	2	1.07	0.26	1.
1	3	1.27	0.64	1.
1	4	1.58	0.90	1.
1	5	1.52	1.02	1.

ABW before merge

min	max	mean	stdd	med
1	2	1.07	0.26	1.
1	1	1.	0	1.
1	3	1.18	0.60	1.
1	2	1.25	0.45	1.
1	3	1.15	0.50	1.

ABW after merge

7 Illustrations

In this section, we provide illustrations of properties highlighted by the statistics investigated. Interface atoms of the two proteins are displayed expanded by a water probe of radius 1.4 Å, while interface water molecules are displayed with their Van der Waals radius —i.e. balls are not expanded to avoid cluttering. The color conventions used to display the atoms and the interface facets are those used in the VMD plugin, and are explained in section 10.

Connectivity, interfaces as a core and a rim.

The purpose of the following examples is to show that (i)structural water fills packing defects and modifies connectivity across the interface (ii)the model of interface as a core and a rim is not general.

Complex 1vfb; Immune system; Fig. 1. This complex, features the Hen Egg-white Lysozyme complexed with the Fv fragments of mouse monoclonal antibody D1.3. The recognition site of Lysozyme is formed by the complementarity determining region (CDR) composed by the combination of the loops of the two variable domains of the heavy and the light chains V_H and V_L of the Fv fragments. The interface atoms of Lysozyme perfectly illustrate the notion of core and rim —Fig. 1(b). But water molecules squeezed in-between the two proteins fill packing defects. These water molecules improve the complementarity between the three chains, and stabilize the whole complex. These molecules result in the interface depicted on Fig. 2(a,b), with two deeps creeks. The same interface without structural water —Fig. 2(c)— has one hole corresponding to a part of one of the two creeks, showing that atoms of the proteins in that regions have a looser packing. It also turns out (see tables) that the interface without water molecules exhibits about twice as many contacts between the atoms.

Complex 1vfb is the only complex where consideration of water molecules increases the connectivity n_g.

Complex 1ao7; Immune system; Fig. 3. This complex features the Tax viral peptide complexed with HLA-A2 and a T Cell Receptor (TCR). The Tax viral peptide is trapped by the HLA-A2 protein in a deep groove formed by three helices and seven beta strands. This Tax/HLA-A2 complex is then presented to the chains α and β of the TCR, which results in a complete burial of the peptide. The interface atoms of these chains —Fig. 3(b) do not follow the usual pattern of core and rim. Notice also that these interface atoms have the topology of a sphere —rather than that of a disk. For this complex, more than one half of the interface atoms are buried. This particularly high value owes to the fact that the viral peptide Tax is entirely trapped.

Complex 1ao7 has the highest average number of neighbors per atom —statistic n_g.

Complex 1nfd; Immune system; Fig. 4. Since on the previous example, the interactions at the level of the TCR are masked, consider Fig. 4 which illustrates a complex between an anti-TCR Fab fragment and a TCR. The TCR consists of the α and β chains. The anti-TCR Fab fragment recognizes a loop localized in the β domain of the TCR. The interface atoms describe a complex geometric pattern, and few atoms are buried.

On the flatness and curvature properties of interfaces.

We now illustrate the curvature properties of interfaces.

Complex 2trc; Signal transduction; Fig. 5. In this complex, the two domains of the phosducin interact with the beta - gamma subunits of transducin, forming a very extensive interface. The interface is mainly due to a vast hydrophobic region near the the N-terminal of phosducin, involving an α-helix of phosducin and seven β strands of the beta subunit of transducin. In particular, three residues (threonine 20P, glycine 21P and proline 22P) form the bottom of a deep pocket in contact with six water molecules. The C-terminal domain interacts with an outer beta sheet of the transducin subunit near the C-terminal of the gamma subunit. Although connected, the interface makes an angle of about 90 degrees in-between the seven β strands and the outer beta sheet of the transducin subunit.

Complex 2trc has the highest s_H statistic.

Complex 1ppe; Proteases; Fig. 6. The interface between the bovine beta-trypsin and Cucurbita maxima trypsin inhibitor I is very characteristic of serine protease / substrate complexes. The active site consists in four different pockets (S1-S4) that recognize four inhibitor residues. In particular, the characteristic S1 binding pocket recognize the side chain of the arginine 5I of the inhibitor in a very deep pocket.

Within proteases, complex 1ppe exhibits the highest fraction of buried atoms in the ABW model.

Multi-patch interfaces and the role of water.

Complex 1tco; Misc complexes; Fig. 7. This complex is a ternary complex involving a fragment of calcineurin A, calcineurin B, the binding protein FKBP12, and the immunosuppressant drug FK506. The interface between the subunits A and B of calcineurin and FKBP12 has the shape of a horse-shoe —enclosing the FK506 drug. More interestingly, this interface has five connected components of significant size, which get reconnected by water molecules.

Complex 1tco is the complex where consideration of water molecules yields the most significant increase in the number of interface atoms.

Complex 1dkg; Misc complexes; Fig. 8. Unlike others multimeric complexes, the dimeric Nucleotide Exchange Factor (GrpE) asymmetrically interacts to one ATPase Domain

of the Molecular Chaperone (DnaK). The interface features three significant patches. The largest one consists of a mainly hydrophobic region localized between the beta sheet domain of GprE and two helices of DnaK on each side of the nucleotide binding site groove. The two smaller patches are located between the long helix of GrpE and the other side of DnaK. Notice the interface atoms forming the second small interface patch on the bottom of Fig. 8 —the first one being hidden by the perspective. Interface atoms shown are those are those of DnaK.

Figure 1: Complex 1vfb (a)Chains: Lysozyme (Grey), antibody Fv fragments (Blue, Red) (b)Interface atoms of the Lysozyme

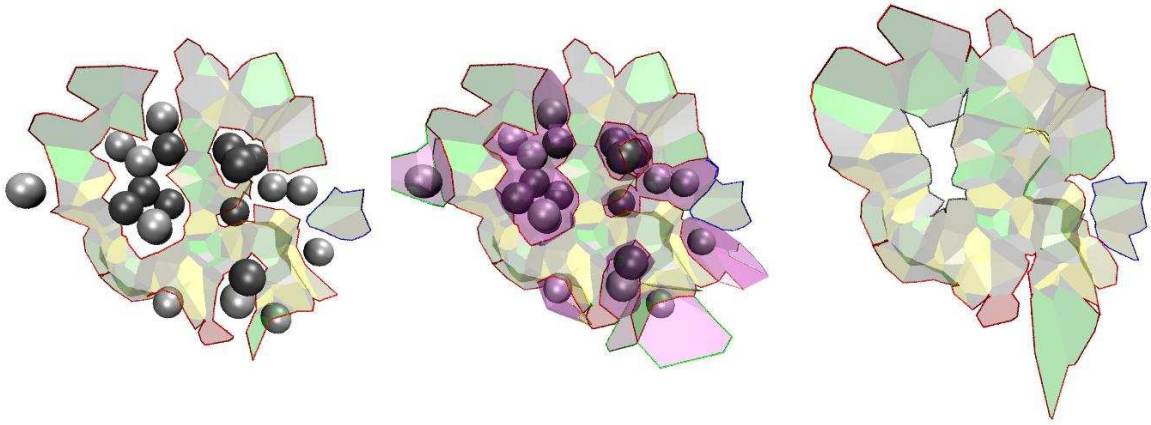


Figure 2: Complex 1vfb (a)Creeks at the interface filled by water molecules (b)Facets of the AW-BW interface shown in purple (c)The interface without water molecules

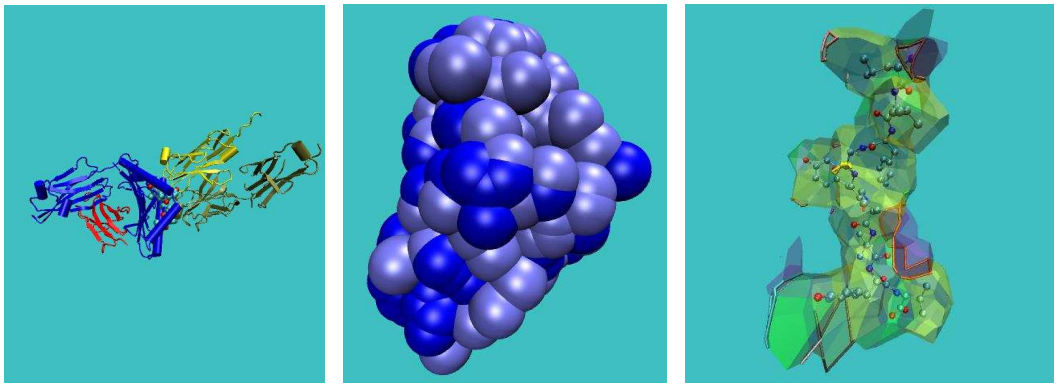
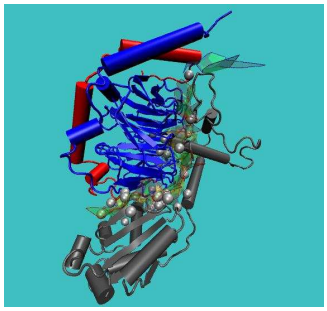


Figure 3: Complex 1ao7 (a)Chains: viral peptide (Van der Waals), HLA-A2 (Blue), TCR chains (α, Yellow; β, Green) (b)Expanded interface atoms of HLA-A2 and TCR chains (c)Viral peptide and interface

Figure 4: Complex 1nfd (a)Chains: TCR (α, Blue; β, Red), anti-TCR Fab fragments in Green and Grey (b)Interface atoms of the TCR chains



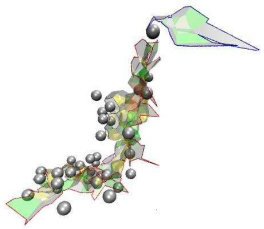
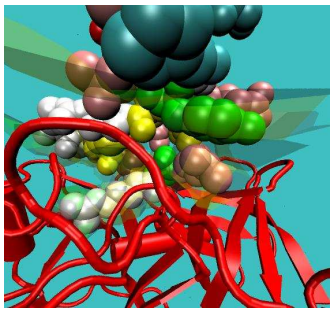


Figure 5: Complex 2trc (a)Chains: Transducin (Blue, Red), Phosducin (Grey) (b)Interface with a bend



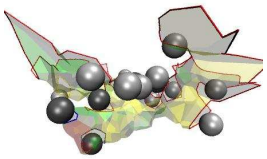


Figure 6: Complex 1ppe (a)Chains: beta-trypsin (Red), Trypsin inhibitor (Colored by residue) (b)Interface with a deep pocket

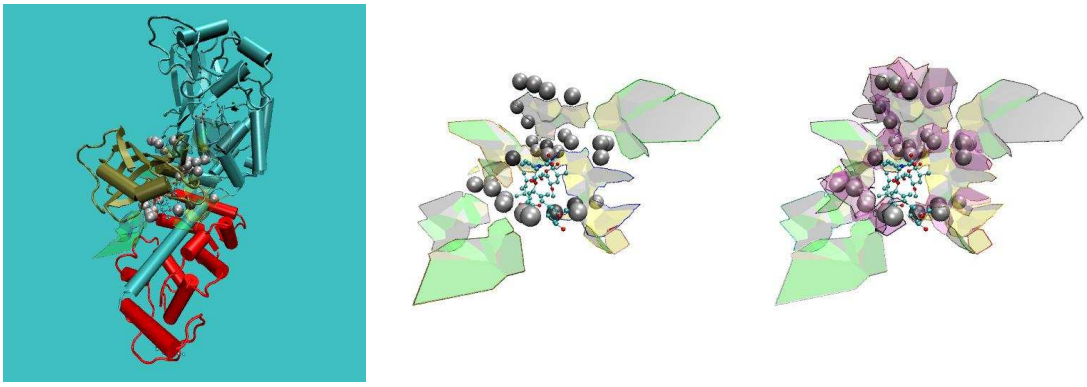


Figure 7: Complex 1tco (a)Chains: Calcineurin A (Cyan), Calcineurin B (Red), FKBP12 (Green), Immuno-suppressant drug FK506 (Van der Waals) (b,c)The AB interface has 5 significant cc, but water molecules bridge them into a single cc

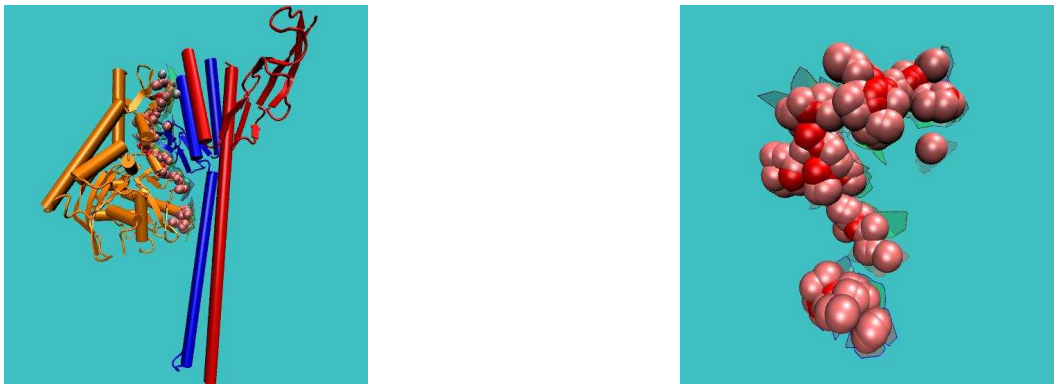


Figure 8: Complex 1dkg. (a)Chains: GrpE (Blue and Red chains), DnaK (Orange) (b)The multi-patch (3 significant patches) structure

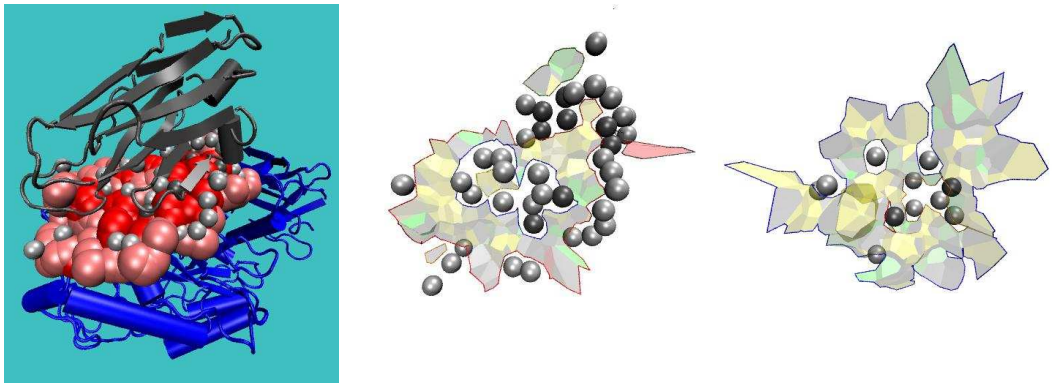


Figure 9: Interfaces with holes (a)1kxv: chains. Interface atoms are from the bottom chain —V_H domain (b)1kxv: one hole with 11 water molecules (c)1udi: three holes with 1,1 and 5 water molecules

8 Appendix: Tables

8.1 Proteases: asymetry of the number of neighbors

Complex	A	B	#AB/#A (AB)	#AB/#B (AB)
1mkw	L (Enzyme)	K (Substrate)	2.83	3.51
3sgb	E (Enzyme)	I (Inhibitor)	3.13	4.01
1brc	I (Inhibitor)	E (Enzyme)	4.74	2.91
1ppf	E (Enzyme)	I (Inhibitor)	3.14	4.00
1tab	I (Inhibitor)	E (Enzyme)	5.36	2.59
4cpa	(Enzyme)	I (Inhibitor)	3.13	3.80
3tpi	Z (Enzyme)	I (Inhibitor)	2.76	4.94
2ptc	I (Inhibitor)	E (Enzyme)	5.18	2.81
2kai	I (Inhibitor)	A (Enzyme)	4.96	2.79
1cbw	A (Enzyme)	D (Inhibitor)	3.17	4.32
1cho	I (Inhibitor)	E (Enzyme)	4.36	3.28
1cse	I (Inhibitor)	E (Enzyme)	4.60	2.90
1mct	A (Enzyme)	I (Inhibitor)	2.84	4.99
1acb	I (Inhibitor)	E (Enzyme)	3.97	3.13
2sic	I (Inhibitor)	E (Enzyme)	4.55	3.11
2sni	I (Inhibitor)	E (Enzyme)	4.18	2.98
1ppe	I (Inhibitor)	E (Enzyme)	4.73	2.99
1tgs	I (Inhibitor)	Z (Enzyme)	4.55	3.05
1avw	B (Inhibitor)	A (Enzyme)	4.73	3.22
1hia	A (Enzyme)	I (Inhibitor)	2.96	4.49
1fle	E (Enzyme)	I (Inhibitor)	3.06	4.05
1stf	I (Inhibitor)	E (Enzyme)	4.30	2.89
1cgi	E (Enzyme)	I (Inhibitor)	3.46	4.11
1bth	P (Inhibitor)	L (Enzyme)	4.11	3.20
4htc	L (Enzyme)	I (Substrate)	3.08	4.30
1tbq	J (Enzyme)	S (Inhibitor)	3.29	4.06
1toc	A (Enzyme)	R (Substrate)	3.23	3.82
1dan	L (Enzyme)	T (Substrate)	3.58	3.15

8.2 Number of atoms and neighbors

PDBId	AB			ABW								AB vs ABW	
	$\#A$	$\#B$	Bur	$\#A$	$\#B$	$\#W$	Bur	$\frac{\#A_{AB}}{\#A}$	$\frac{\#A_{AW}}{\#A}$	$\frac{\#B_{AB}}{\#B}$	$\frac{\#B_{BW}}{\#B}$	$R(\#A+B)$	$R(bur)$
Protease													
1mkw	93	75	0.17	100	113	9	0.26	0.89	0.40	0.65	0.62	1.27	1.58
3sgb	95	74	0.43	128	96	14	0.53	0.72	0.55	0.72	0.60	1.33	1.24
1brc	62	101	0.44	66	122	3	0.48	0.92	0.15	0.82	0.32	1.15	1.10
1ppf	98	77	0.48	151	106	14	0.56	0.64	0.58	0.73	0.63	1.47	1.18
1tab	58	120	0.42	59	143	5	0.48	0.98	0.27	0.79	0.41	1.13	1.16
4cpa	79	65	0.44	79	65	0	0.44	1.00	0.00	1.00	0.00	1.00	1.00
3tpi	120	67	0.48	172	85	12	0.58	0.67	0.55	0.79	0.47	1.37	1.22
2ptc	65	120	0.45	81	163	13	0.51	0.80	0.51	0.73	0.49	1.32	1.13
2kai	68	121	0.47	68	134	2	0.48	1.00	0.09	0.90	0.16	1.07	1.02
1cbw	101	74	0.46	162	87	11	0.52	0.62	0.65	0.85	0.49	1.42	1.12
1cho	76	101	0.42	114	159	25	0.54	0.64	0.75	0.61	0.72	1.54	1.29
1cse	72	114	0.48	92	142	11	0.59	0.73	0.62	0.79	0.48	1.26	1.24
1mct	130	74	0.42	172	87	12	0.63	0.72	0.53	0.83	0.54	1.27	1.48
1acb	86	109	0.38	120	143	12	0.44	0.69	0.57	0.74	0.49	1.35	1.15
2sic	84	123	0.47	116	163	14	0.59	0.71	0.63	0.72	0.60	1.35	1.24
2sni	87	122	0.40	110	149	8	0.47	0.76	0.53	0.81	0.40	1.24	1.16
1ppe	88	139	0.44	107	191	17	0.63	0.78	0.62	0.66	0.64	1.31	1.45
1tgs	88	131	0.41	91	150	6	0.47	0.97	0.24	0.85	0.40	1.10	1.16
1avw	94	138	0.49	100	189	10	0.62	0.93	0.34	0.68	0.56	1.25	1.27
1hia	123	81	0.39	179	110	18	0.52	0.66	0.68	0.72	0.74	1.42	1.33
1fle	122	92	0.37	138	96	5	0.44	0.88	0.31	0.96	0.30	1.09	1.19
1stf	90	134	0.37	138	172	21	0.55	0.61	0.80	0.75	0.57	1.38	1.49
1cgi	145	122	0.51	173	126	3	0.55	0.82	0.27	0.97	0.09	1.12	1.07
1bth	130	167	0.43	139	208	6	0.47	0.94	0.18	0.78	0.37	1.17	1.09
4htc	226	162	0.28	308	194	26	0.44	0.70	0.65	0.82	0.64	1.29	1.58
1tbq	235	190	0.32	289	192	7	0.36	0.81	0.33	0.99	0.15	1.13	1.15
1toc	233	197	0.31	233	197	0	0.31	1.00	0.00	1.00	0.00	1.00	1.00
1dan	198	225	0.20	293	295	39	0.44	0.63	0.70	0.73	0.60	1.39	2.17
Immune system													
1wej	73	90	0.39	126	128	21	0.56	0.56	0.77	0.64	0.65	1.56	1.43
1jhl	77	89	0.30	77	89	0	0.30	1.00	0.00	1.00	0.00	1.00	1.00
2vir	91	74	0.24	91	74	0	0.24	1.00	0.00	1.00	0.00	1.00	1.00
2jel	81	105	0.33	109	121	9	0.56	0.72	0.61	0.82	0.48	1.24	1.68
1bvk	77	86	0.21	77	86	0	0.21	1.00	0.00	1.00	0.00	1.00	1.00
1vfb	90	84	0.24	133	116	21	0.54	0.63	0.80	0.66	0.86	1.43	2.28
1mlc	82	98	0.42	98	114	7	0.57	0.84	0.41	0.82	0.42	1.18	1.34
1nmb	79	91	0.24	89	107	5	0.31	0.89	0.30	0.81	0.41	1.15	1.30
1osp	87	95	0.32	154	140	30	0.52	0.50	0.83	0.61	0.74	1.62	1.63
1eo8	101	91	0.32	113	106	5	0.42	0.87	0.32	0.85	0.31	1.14	1.34
3hfm	105	100	0.42	105	100	0	0.42	1.00	0.00	1.00	0.00	1.00	1.00
1kxt	102	101	0.28	148	138	20	0.39	0.62	0.75	0.72	0.65	1.41	1.38
1kxv	94	110	0.25	149	173	39	0.60	0.56	0.87	0.53	0.87	1.58	2.46
1bql	94	102	0.24	108	134	9	0.41	0.87	0.46	0.75	0.60	1.23	1.69
1dvf	108	92	0.34	145	137	15	0.49	0.74	0.62	0.66	0.71	1.41	1.46
1mel	101	86	0.37	103	86	1	0.40	0.97	0.13	1.00	0.06	1.01	1.08
1nfd	84	119	0.33	84	119	0	0.33	1.00	0.00	1.00	0.00	1.00	1.00
1fbi	119	102	0.35	119	102	0	0.35	1.00	0.00	1.00	0.00	1.00	1.00
3hfl	110	97	0.27	125	104	4	0.33	0.86	0.25	0.93	0.26	1.11	1.23
1dqj	114	121	0.45	156	162	20	0.56	0.69	0.67	0.71	0.60	1.35	1.25
1nsn	100	107	0.17	100	107	0	0.17	1.00	0.00	1.00	0.00	1.00	1.00
1qfu	116	106	0.26	135	121	8	0.43	0.84	0.41	0.86	0.43	1.15	1.64
1ahw	114	129	0.32	114	129	0	0.32	1.00	0.00	1.00	0.00	1.00	1.00
1iai	112	120	0.27	112	120	0	0.27	1.00	0.00	1.00	0.00	1.00	1.00
1nca	116	132	0.37	123	132	1	0.37	0.94	0.09	1.00	0.02	1.03	0.99
1ao7	208	75	0.51	229	75	4	0.56	0.91	0.21	1.00	0.25	1.07	1.10
1kxq	131	138	0.31	194	212	39	0.58	0.62	0.76	0.59	0.75	1.51	1.88
1kb5	149	138	0.26	189	159	11	0.34	0.77	0.48	0.86	0.36	1.21	1.30
Enzyme Complexes													
2pcc	55	63	0.14	85	90	13	0.27	0.62	0.74	0.68	0.78	1.48	1.90
1gla	80	71	0.21	80	71	0	0.21	1.00	0.00	1.00	0.00	1.00	1.00
2tec	85	116	0.47	108	144	9	0.56	0.73	0.64	0.79	0.46	1.25	1.20
1brs	99	108	0.29	145	150	22	0.52	0.64	0.81	0.61	0.79	1.43	1.75

1fss	113	136	0.36	117	155	4	0.35	0.96	0.15	0.88	0.21	1.09	0.98
1ydr	133	96	0.28	158	101	9	0.39	0.83	0.36	0.94	0.34	1.13	1.41
1udi	118	115	0.30	124	160	11	0.44	0.95	0.42	0.68	0.61	1.22	1.44
1mah	121	143	0.32	121	143	0	0.32	1.00	0.00	1.00	0.00	1.00	1.00
1ugh	122	122	0.33	131	181	17	0.42	0.92	0.38	0.64	0.66	1.28	1.29
1dfj	155	159	0.20	169	182	7	0.25	0.89	0.28	0.85	0.30	1.12	1.24
1dhk	194	182	0.31	266	207	31	0.60	0.68	0.71	0.85	0.61	1.26	1.92
Signal transduction													
1a0o	73	65	0.24	73	65	0	0.24	1.00	0.00	1.00	0.00	1.00	1.00
1gua	79	77	0.32	102	86	9	0.45	0.75	0.55	0.87	0.43	1.21	1.39
1a2k	99	95	0.28	149	112	13	0.46	0.61	0.66	0.83	0.51	1.35	1.64
1agr	97	107	0.31	131	116	8	0.39	0.72	0.48	0.91	0.28	1.21	1.24
1tx4	130	130	0.35	200	208	51	0.68	0.57	0.79	0.59	0.80	1.57	1.93
1gg2	141	144	0.34	183	213	29	0.40	0.75	0.50	0.66	0.59	1.39	1.16
1got	149	159	0.35	238	191	27	0.48	0.59	0.74	0.81	0.58	1.39	1.37
1aip	167	159	0.27	173	166	2	0.27	0.97	0.08	0.96	0.10	1.04	1.02
1fin	194	222	0.34	241	267	21	0.45	0.79	0.51	0.81	0.48	1.22	1.32
1efu	206	234	0.33	298	331	65	0.51	0.65	0.69	0.66	0.69	1.43	1.54
3hhr	255	267	0.33	255	267	0	0.33	1.00	0.00	1.00	0.00	1.00	1.00
2trc	290	272	0.36	393	352	37	0.43	0.73	0.54	0.76	0.56	1.33	1.22
Misc													
1l0y	63	74	0.18	110	106	16	0.28	0.56	0.77	0.68	0.72	1.58	1.61
1ak4	76	51	0.31	92	53	4	0.35	0.80	0.45	0.96	0.26	1.14	1.12
1avz	74	78	0.30	74	78	0	0.30	1.00	0.00	1.00	0.00	1.00	1.00
1efn	78	79	0.42	93	83	4	0.45	0.83	0.40	0.95	0.25	1.12	1.08
1fc2	76	87	0.38	76	87	0	0.38	1.00	0.00	1.00	0.00	1.00	1.00
1seb	79	85	0.35	79	85	0	0.35	1.00	0.00	1.00	0.00	1.00	1.00
1igc	79	82	0.37	112	108	14	0.42	0.70	0.70	0.76	0.61	1.37	1.14
2mta	81	87	0.25	98	119	10	0.44	0.79	0.59	0.73	0.59	1.29	1.75
1ycs	89	100	0.32	145	124	15	0.45	0.60	0.72	0.81	0.55	1.42	1.41
1kkl	100	110	0.34	112	142	8	0.34	0.88	0.28	0.76	0.39	1.21	1.01
1atn	120	101	0.32	120	101	0	0.32	1.00	0.00	1.00	0.00	1.00	1.00
1fq1	102	103	0.22	102	103	0	0.22	1.00	0.00	1.00	0.00	1.00	1.00
1ebp	112	123	0.38	112	123	0	0.38	1.00	0.00	1.00	0.00	1.00	1.00
1dkg	118	115	0.24	134	127	4	0.25	0.88	0.21	0.91	0.20	1.12	1.03
2btf	130	122	0.34	130	122	0	0.34	1.00	0.00	1.00	0.00	1.00	1.00
1spb	120	168	0.40	128	215	19	0.52	0.91	0.55	0.74	0.58	1.19	1.32
1tco	106	104	0.19	191	163	28	0.41	0.53	0.74	0.60	0.75	1.69	2.22
1wq1	147	178	0.30	167	199	10	0.39	0.86	0.42	0.89	0.31	1.13	1.32
1hwg	264	263	0.34	305	331	21	0.41	0.85	0.38	0.77	0.48	1.21	1.20

8.3 Connectivity

PDBId	*AB*				*ABW*				*AB* vs *ABW*
	n_A	n_B	r_{Mm}	n_g	n_A	n_B	r_{Mm}	n_g	$R(n_g)$
Protease									
1mkw	2.83	3.51	1.24	3.13	2.91	2.87	1.01	2.89	0.92
3sgb	3.13	4.01	1.28	3.51	2.95	3.72	1.26	3.28	0.93
1brc	4.74	2.91	1.63	3.61	4.53	2.69	1.69	3.34	0.92
1ppf	3.14	4.00	1.27	3.52	2.76	3.60	1.30	3.11	0.88
1tab	5.36	2.59	2.07	3.49	5.32	2.50	2.13	3.32	0.95
4cpa	3.13	3.80	1.22	3.43	3.13	3.80	1.22	3.43	1.00
3tpi	2.76	4.94	1.79	3.54	2.54	4.40	1.73	3.16	0.89
2ptc	5.18	2.81	1.85	3.64	4.70	2.67	1.76	3.34	0.92
2kai	4.96	2.79	1.78	3.57	5.03	2.67	1.88	3.47	0.97
1cbw	3.17	4.32	1.36	3.66	2.62	4.15	1.59	3.15	0.86
1cho	4.36	3.28	1.33	3.74	3.87	2.91	1.33	3.31	0.89
1cse	4.60	2.90	1.58	3.56	4.08	2.79	1.46	3.29	0.93
1mct	2.84	4.99	1.76	3.62	2.69	4.72	1.75	3.37	0.93
1acb	3.97	3.13	1.27	3.50	3.48	2.85	1.22	3.13	0.90
2sic	4.55	3.11	1.46	3.69	3.97	2.98	1.33	3.39	0.92
2sni	4.18	2.98	1.40	3.48	3.79	2.79	1.36	3.22	0.92
1ppe	4.73	2.99	1.58	3.67	4.46	2.84	1.57	3.42	0.93
1tgs	4.55	3.05	1.49	3.65	4.55	3.01	1.51	3.59	0.98
1avw	4.73	3.22	1.47	3.84	4.67	2.90	1.61	3.52	0.92
1hia	2.96	4.49	1.52	3.57	2.80	4.06	1.45	3.28	0.92
1fle	3.06	4.05	1.33	3.49	2.99	4.16	1.39	3.47	0.99
1stf	4.30	2.89	1.49	3.46	3.75	2.89	1.30	3.27	0.95
1cgi	3.46	4.11	1.19	3.75	3.15	4.04	1.28	3.53	0.94
1bth	4.11	3.20	1.28	3.60	3.96	2.90	1.37	3.33	0.93
4htc	3.08	4.30	1.40	3.59	2.92	4.14	1.42	3.39	0.95
1tbq	3.29	4.06	1.24	3.63	2.98	4.14	1.39	3.44	0.95
1toc	3.23	3.82	1.18	3.50	3.23	3.82	1.18	3.50	1.00
1dan	3.58	3.15	1.14	3.35	3.29	3.13	1.05	3.21	0.96
Immune system									
1wej	3.93	3.19	1.23	3.52	3.25	3.05	1.07	3.15	0.90
1jhl	3.79	3.28	1.16	3.52	3.79	3.28	1.16	3.52	1.00

2vir	3.00	3.69	1.23	3.31	3.00	3.69	1.23	3.31	1.00
2jel	4.07	3.14	1.30	3.55	3.61	3.12	1.16	3.36	0.95
1bvk	3.64	3.26	1.12	3.44	3.64	3.26	1.12	3.44	1.00
1vfb	3.12	3.35	1.07	3.23	3.17	3.53	1.11	3.33	1.03
1mlc	3.94	3.30	1.20	3.59	3.68	3.22	1.14	3.43	0.96
1nmb	3.53	3.07	1.15	3.28	3.33	2.93	1.14	3.11	0.95
1osp	3.53	3.23	1.09	3.37	3.08	3.19	1.04	3.13	0.93
1eo8	3.20	3.55	1.11	3.36	3.13	3.26	1.04	3.20	0.95
3hfm	3.82	4.01	1.05	3.91	3.82	4.01	1.05	3.91	1.00
1kxt	3.52	3.55	1.01	3.54	3.16	3.20	1.02	3.18	0.90
1kxv	3.39	2.90	1.17	3.13	3.29	2.98	1.10	3.12	1.00
1bql	3.65	3.36	1.09	3.50	3.54	3.14	1.13	3.32	0.95
1dvf	3.16	3.71	1.17	3.41	3.04	3.28	1.08	3.16	0.93
1mel	3.23	3.79	1.17	3.49	3.26	3.81	1.17	3.51	1.01
1nfd	4.27	3.02	1.42	3.54	4.27	3.02	1.42	3.54	1.00
1fbi	3.34	3.89	1.17	3.59	3.34	3.89	1.17	3.59	1.00
3hfl	3.30	3.74	1.13	3.51	3.18	3.72	1.17	3.42	0.98
1dqj	3.77	3.55	1.06	3.66	3.51	3.35	1.05	3.42	0.94
1nsn	3.29	3.07	1.07	3.18	3.29	3.07	1.07	3.18	1.00
1qfu	3.22	3.52	1.09	3.36	2.99	3.36	1.13	3.16	0.94
1ahw	3.76	3.33	1.13	3.53	3.76	3.33	1.13	3.53	1.00
1iai	3.50	3.27	1.07	3.38	3.50	3.27	1.07	3.38	1.00
1nca	3.62	3.18	1.14	3.39	3.50	3.20	1.10	3.35	0.99
1ao7	2.77	7.69	2.77	4.08	2.72	7.88	2.90	3.99	0.98
1kxq	3.42	3.25	1.05	3.33	3.27	3.03	1.08	3.15	0.94
1kb5	3.17	3.42	1.08	3.29	2.98	3.43	1.15	3.19	0.97
Enzyme Complexes									
2pcc	3.24	2.83	1.15	3.02	2.95	2.84	1.04	2.90	0.96
1gla	2.91	3.28	1.13	3.09	2.91	3.28	1.13	3.09	1.00
2tec	4.16	3.05	1.36	3.52	3.80	2.80	1.36	3.23	0.92
1brs	3.76	3.44	1.09	3.59	3.30	3.25	1.02	3.27	0.91
1fss	3.90	3.24	1.20	3.54	3.92	3.05	1.29	3.43	0.97
1ydr	3.14	4.35	1.39	3.65	3.03	4.51	1.49	3.61	0.99
1udi	3.25	3.34	1.03	3.30	3.42	3.08	1.11	3.23	0.98
1mah	3.86	3.27	1.18	3.54	3.86	3.27	1.18	3.54	1.00
1ugh	3.31	3.31	1.00	3.31	3.49	3.07	1.14	3.25	0.98
1dfj	3.19	3.11	1.03	3.15	3.16	2.97	1.06	3.06	0.97

1dhk	3.46	3.69	1.07	3.57	3.35	3.90	1.16	3.59	1.01
Signal transduction									
1a0o	2.95	3.31	1.12	3.12	2.95	3.31	1.12	3.12	1.00
1gua	3.34	3.43	1.03	3.38	3.26	3.50	1.07	3.37	1.00
1a2k	3.40	3.55	1.04	3.47	2.91	3.46	1.19	3.14	0.90
1agr	3.33	3.02	1.10	3.17	2.97	3.04	1.02	3.00	0.95
1tx4	3.45	3.45	1.00	3.45	3.48	3.34	1.04	3.40	0.99
1gg2	3.31	3.24	1.02	3.28	3.21	2.95	1.09	3.07	0.94
1got	3.30	3.09	1.07	3.19	2.91	3.19	1.10	3.04	0.95
1aip	3.32	3.49	1.05	3.40	3.28	3.44	1.05	3.36	0.99
1fin	3.72	3.25	1.14	3.47	3.53	3.21	1.10	3.36	0.97
1efu	3.54	3.12	1.14	3.31	3.33	3.14	1.06	3.23	0.97
3hhr	3.67	3.51	1.05	3.59	3.67	3.51	1.05	3.59	1.00
2trc	3.39	3.62	1.07	3.50	3.11	3.39	1.09	3.24	0.92
Misc									
1l0y	3.40	2.89	1.17	3.12	2.79	2.92	1.05	2.86	0.91
1ak4	2.78	4.14	1.49	3.32	2.68	4.13	1.54	3.21	0.97
1avz	3.77	3.58	1.05	3.67	3.77	3.58	1.05	3.67	1.00
1efn	3.73	3.68	1.01	3.71	3.49	3.69	1.05	3.59	0.97
1fc2	3.99	3.48	1.14	3.72	3.99	3.48	1.14	3.72	1.00
1seb	3.49	3.25	1.08	3.37	3.49	3.25	1.08	3.37	1.00
1igc	3.48	3.35	1.04	3.42	3.25	3.25	1.00	3.25	0.95
2mta	3.44	3.21	1.07	3.32	3.34	2.87	1.16	3.08	0.93
1ycs	3.73	3.32	1.12	3.51	3.11	3.22	1.03	3.16	0.90
1kkl	3.33	3.03	1.10	3.17	3.24	2.75	1.18	2.96	0.93
1atn	3.23	3.83	1.19	3.50	3.23	3.83	1.19	3.50	1.00
1fq1	3.50	3.47	1.01	3.48	3.50	3.47	1.01	3.48	1.00
1ebp	3.79	3.46	1.10	3.62	3.79	3.46	1.10	3.62	1.00
1dkg	2.92	3.00	1.03	2.96	2.79	2.91	1.04	2.85	0.96
2btf	3.50	3.73	1.07	3.61	3.50	3.73	1.07	3.61	1.00
1spb	4.33	3.09	1.40	3.60	4.50	2.96	1.52	3.53	0.98
1tco	2.89	2.94	1.02	2.91	2.45	2.73	1.12	2.58	0.88
1wq1	3.88	3.21	1.21	3.51	3.85	3.20	1.20	3.49	0.99
1hwg	3.45	3.46	1.00	3.45	3.35	3.24	1.03	3.29	0.95

8.4 Surface area

PDB Id	Area values					Comparison			
	$A_{Ref.}$	A_{BER}	A_{AB}	A_{ABW}	A_{ABW-w}	$\frac{A_{BER}}{A_{Ref.}}$	$\frac{A_{AB}}{A_{Ref.}}$	$\frac{A_{ABW}}{A_{Ref.}}$	$\frac{A_{ABW-w}}{A_{ABW}}$
Protease									
1mkw	640		1071.73	1286.31	876.53		1.67	2.01	0.68
3sgb	640	462	857.41	1055.47	682.09	0.72	1.34	1.65	0.65
1brc	660		831.70	907.73	773.21		1.26	1.38	0.85
1ppf	665	900	865.33	1173.68	677.26	1.35	1.30	1.76	0.58
1tab	680		854.35	918.10	762.37		1.26	1.35	0.83
4cpa	680	672	779.59	779.59	779.59	0.99	1.15	1.15	1.00
3tpi	710	643	892.91	1189.08	756.78	0.91	1.26	1.67	0.64
2ptc	715	575	916.31	1237.04	756.75	0.80	1.28	1.73	0.61
2kai	720	776	899.91	977.71	880.29	1.08	1.25	1.36	0.90
1cbw	730	653	860.87	1196.25	706.26	0.89	1.18	1.64	0.59
1cho	735	736	982.75	1406.01	654.23	1.00	1.34	1.91	0.47
1cse	745	745	976.05	1112.11	819.28	1.00	1.31	1.49	0.74
1mct	760	694	944.92	1077.97	748.86	0.91	1.24	1.42	0.69
1acb	770	717	1003.68	1232.23	855.31	0.93	1.30	1.60	0.69
2sic	810	717	1020.79	1267.59	828.28	0.89	1.26	1.56	0.65
2sni	815	869	1186.24	1284.02	997.62	1.07	1.46	1.58	0.78
1ppe	845		1047.04	1339.95	843.35		1.24	1.59	0.63
1tgs	865	734	1149.33	1265.02	1045.34	0.85	1.33	1.46	0.83
1avw	870	1011	1043.05	1253.57	918.02	1.16	1.20	1.44	0.73
1hia	870	847	1077.68	1412.02	825.78	0.97	1.24	1.62	0.58
1fle	890	546	1223.43	1314.58	1124.30	0.61	1.37	1.48	0.86
1stf	895	718	1196.45	1532.52	928.25	0.80	1.34	1.71	0.61
1cgi	1025		1337.26	1401.51	1297.92		1.30	1.37	0.93
1bth	1190	872	1442.70	1531.14	1360.72	0.73	1.21	1.29	0.89
4htc	1675	1035	2261.21	2662.51	1780.82	0.62	1.35	1.59	0.67
1tbq	1755	1477	2271.12	2433.23	2177.33	0.84	1.29	1.39	0.89
1toc	1755	1386	2244.35	2244.35	2244.35	0.79	1.28	1.28	1.00
1dan	1885	1859	2620.30	3171.63	2108.52	0.99	1.39	1.68	0.66
Immune system complexes									
1wej	590		900.72	1222.57	655.72		1.53	2.07	0.54
1jhl	630	638	860.47	860.47	860.47	1.01	1.37	1.37	1.00

2vir	630		848.25	848.25	848.25		1.35	1.35	1.00
2jel	680	638	962.03	1057.65	796.06	0.94	1.41	1.56	0.75
1bvk	700		901.80	901.80	901.80		1.29	1.29	1.00
1vfb	700	585	944.05	1159.11	627.39	0.84	1.35	1.66	0.54
1mlc	705	510	954.87	1088.42	865.57	0.72	1.35	1.54	0.80
1nmb	750	921	903.51	1022.30	809.19	1.23	1.20	1.36	0.79
1osp	750	747	931.25	1403.62	610.34	1.00	1.24	1.87	0.43
1eo8	765		1014.12	1077.06	927.88		1.33	1.41	0.86
3hfm	805	825	1093.88	1093.88	1093.88	1.02	1.36	1.36	1.00
1kxt	810		1023.02	1314.24	744.13		1.26	1.62	0.57
1kxv	810		1040.36	1477.22	605.99		1.28	1.82	0.41
1bql	815		1214.65	1316.36	1064.33		1.49	1.62	0.81
1dvf	840	775	1079.56	1428.19	901.57	0.92	1.29	1.70	0.63
1mel	855	502	962.61	975.42	947.05	0.59	1.13	1.14	0.97
1nfd	855	904	1073.22	1073.22	1073.22	1.06	1.26	1.26	1.00
1fbi	860	617	1157.06	1157.06	1157.06	0.72	1.35	1.35	1.00
3hfl	865	719	1232.84	1297.57	1181.26	0.83	1.43	1.50	0.91
1dqj	880		1242.94	1606.81	988.98		1.41	1.83	0.62
1nsn	900	1089	1260.44	1260.44	1260.44	1.21	1.40	1.40	1.00
1qfu	920	1307	1246.57	1340.67	1119.99	1.42	1.35	1.46	0.84
1ahw	950		1203.12	1203.12	1203.12		1.27	1.27	1.00
1iai	950	1000	1233.97	1233.97	1233.97	1.05	1.30	1.30	1.00
1nca	980	1308	1312.00	1345.02	1293.33	1.33	1.34	1.37	0.96
1ao7	995	866	1236.97	1300.94	1188.33	0.87	1.24	1.31	0.91
1kxq	1070		1398.51	1968.68	987.95		1.31	1.84	0.50
1kb5	1170	1151	1650.53	1844.26	1475.27	0.98	1.41	1.58	0.80
Enzyme Complexes									
2pcc	585	580	759.98	968.78	584.47	0.99	1.30	1.66	0.60
1gla	650	712	914.85	914.85	914.85	1.10	1.41	1.41	1.00
2tec	780		963.18	1075.50	802.82		1.23	1.38	0.75
1brs	785	703	998.06	1387.41	690.20	0.90	1.27	1.77	0.50
1fss	985	728	1373.62	1503.85	1277.87	0.74	1.39	1.53	0.85
1ydr	1000	783	1311.82	1408.88	1218.86	0.78	1.31	1.41	0.87
1udi	1010	906	1220.56	1365.64	1061.72	0.90	1.21	1.35	0.78
1mah	1075		1498.98	1498.98	1498.98		1.39	1.39	1.00
1ugh	1095		1401.55	1676.21	1201.13		1.28	1.53	0.72
1dfj	1300	1795	1866.42	2041.44	1711.12	1.38	1.44	1.57	0.84

1dhk	1540	1686	1810.42	2138.37	1436.20	1.09	1.18	1.39	0.67
Signal transduction									
1a0o	570	397	863.05	863.05	863.05	0.70	1.51	1.51	1.00
1gua	655	617	926.14	1025.73	808.18	0.94	1.41	1.57	0.79
1a2k	795	966	973.70	1189.44	803.38	1.22	1.22	1.50	0.68
1agr	825	1278	1093.01	1246.64	1025.00	1.55	1.32	1.51	0.82
1tx4	1140	1219	1472.49	2010.84	979.31	1.07	1.29	1.76	0.49
1gg2	1180	1778	1728.14	2185.05	1341.46	1.51	1.46	1.85	0.61
1got	1250	1550	1746.50	2102.99	1311.00	1.24	1.40	1.68	0.62
1aip	1470	1639	1898.23	1932.72	1842.38	1.11	1.29	1.31	0.95
1fin	1700	1533	2258.37	2535.75	1899.65	0.90	1.33	1.49	0.75
1efu	1830	2205	2484.64	3238.34	1700.25	1.20	1.36	1.77	0.53
3hhr	2075		2716.60	2716.60	2716.60		1.31	1.31	1.00
2trc	2330	2408	2959.13	3634.20	2422.90	1.03	1.27	1.56	0.67
Other complexes									
1l0y	565		805.79	1118.41	574.70		1.43	1.98	0.51
1ak4	585	409	724.62	796.14	655.29	0.70	1.24	1.36	0.82
1avz	630		858.35	858.35	858.35		1.36	1.36	1.00
1efn	630	488	906.98	964.16	808.08	0.77	1.44	1.53	0.84
1fc2	650	604	881.16	881.16	881.16	0.93	1.36	1.36	1.00
1seb	670	1081	874.19	874.19	874.19	1.61	1.30	1.30	1.00
1igc	675	498	876.39	1281.58	666.83	0.74	1.30	1.90	0.52
2mta	730		987.88	1140.72	851.53		1.35	1.56	0.75
1ycs	750	560	1196.63	1343.89	834.57	0.75	1.60	1.79	0.62
1kkl	820		1175.19	1445.53	1058.23		1.43	1.76	0.73
1atn	890	796	1177.32	1177.32	1177.32	0.89	1.32	1.32	1.00
1fq1	915		1407.35	1407.35	1407.35		1.54	1.54	1.00
1ebp	970		1245.52	1245.52	1245.52		1.28	1.28	1.00
1dkg	990	1662	1422.20	1514.74	1357.34	1.68	1.44	1.53	0.90
2btf	1045	1048	1382.88	1382.88	1382.88	1.00	1.32	1.32	1.00
1spb	1115		1627.81	1841.12	1120.84		1.46	1.65	0.61
1tco	1235		1401.50	1858.18	1102.19		1.13	1.50	0.59
1wq1	1455		1804.65	1947.63	1669.06		1.24	1.34	0.86
1hwg	2100	2022	2913.88	3370.14	2610.34	0.96	1.39	1.60	0.77

8.5 Connected components

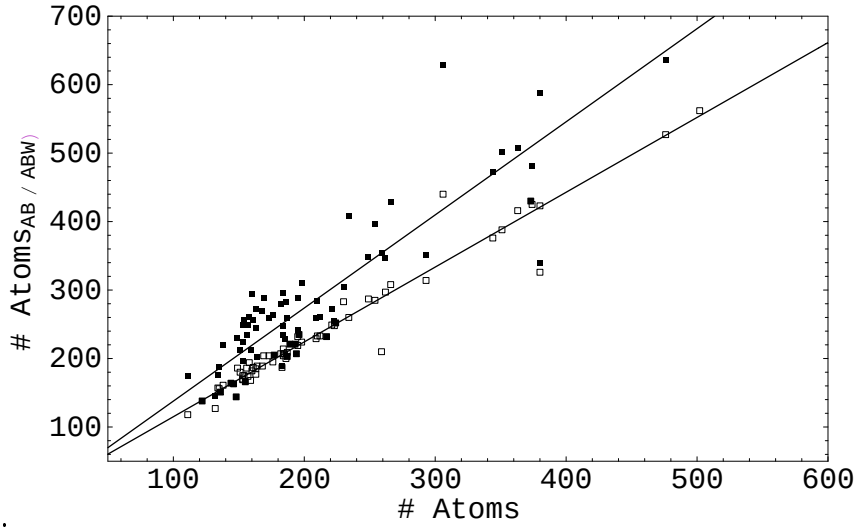
PDBId	connected components					
	AB		*ABW*			
	$\#cc$	$\#scc$	$\#cc_{bm}$	$\#scc_{bm}$	$\#cc_{am}$	$\#scc_{am}$
Protease						
1mkw	2	2	2	2	2	2
3sgb	1	1	2	1	2	1
1brc	1	1	1	1	1	1
1ppf	3	1	5	1	1	1
1tab	1	1	1	1	1	1
4cpa	1	1	1	1	1	1
3tpi	1	1	1	1	1	1
2ptc	1	1	1	1	1	1
2kai	2	1	2	1	2	1
1cbw	1	1	1	1	1	1
1cho	2	1	2	1	2	1
1cse	2	1	3	2	1	1
1mct	1	1	2	1	1	1
1acb	1	1	2	1	3	1
2sic	1	1	1	1	1	1
2sni	2	1	3	1	2	1
1ppe	1	1	1	1	1	1
1tgs	2	1	2	1	2	1
1avw	2	1	2	1	1	1
1hia	1	1	1	1	1	1
1fle	2	1	2	1	1	1
1stf	2	1	2	1	1	1
1cgi	1	1	1	1	1	1
1bth	1	1	1	1	1	1
4htc	1	1	2	1	1	1
1tbq	2	2	3	2	2	2
1toc	2	1	2	1	2	1
1dan	4	3	6	3	2	1
Immune system						

1wej	3	2	3	2	1	1
1jhl	1	1	1	1	1	1
2vir	1	1	1	1	1	1
2jel	1	1	1	1	1	1
1bvk	2	1	2	1	2	1
1vfb	2	1	2	1	1	1
1mlc	1	1	1	1	1	1
1nmb	2	1	2	1	2	1
1osp	2	2	3	2	1	1
1eo8	1	1	1	1	1	1
3hfm	1	1	1	1	1	1
1kxt	1	1	1	1	1	1
1kxv	2	1	6	1	1	1
1bql	1	1	1	1	1	1
1dvf	3	1	3	1	1	1
1mel	2	1	2	1	2	1
1nfd	2	1	2	1	2	1
1fbi	1	1	1	1	1	1
3hfl	1	1	1	1	1	1
1dqj	1	1	1	1	1	1
1nsn	4	1	4	1	4	1
1qfu	1	1	1	1	1	1
1ahw	4	1	4	1	4	1
1iai	1	1	1	1	1	1
1nca	4	1	4	1	4	1
1ao7	1	1	1	1	1	1
1kxq	3	1	3	1	2	1
1kb5	2	1	2	1	2	1
Enzyme Complexes						
2pcc	2	2	2	2	1	1
1gla	2	1	2	1	2	1
2tec	3	1	4	1	1	1
1brs	1	1	2	1	1	1
1fss	2	1	2	1	1	1
1ydr	2	1	2	1	1	1
1udi	1	1	1	1	1	1
1mah	1	1	1	1	1	1

1ugh	2	1	2	1	1	1
1dfj	5	3	5	3	5	3
1dhk	3	1	3	1	2	1
Signal transduction						
1a0o	2	1	2	1	2	1
1gua	1	1	1	1	1	1
1a2k	1	1	1	1	1	1
1agr	2	1	2	1	2	1
1tx4	3	1	4	2	2	1
1gg2	4	2	4	2	3	1
1got	4	2	4	2	2	2
1aip	2	2	2	2	2	2
1fin	1	1	1	1	1	1
1efu	5	3	5	4	3	2
3hhr	3	1	3	1	3	1
2trc	3	1	3	1	1	1
Misc						
1l0y	2	2	2	2	1	1
1ak4	1	1	1	1	1	1
1avz	1	1	1	1	1	1
1efn	1	1	1	1	1	1
1fc2	1	1	1	1	1	1
1seb	1	1	1	1	1	1
1igc	2	1	2	1	1	1
2mta	1	1	1	1	1	1
1ycs	4	2	4	2	1	1
1kkl	4	1	4	1	3	1
1atn	1	1	1	1	1	1
1fq1	3	2	3	2	3	2
1ebp	1	1	1	1	1	1
1dkg	5	3	5	3	4	3
2btf	1	1	1	1	1	1
1spb	1	1	1	1	1	1
1tco	6	2	9	5	2	1
1wq1	5	1	5	1	5	1
1hwg	3	1	3	2	1	1

9 Figures

This section provides plots for several statistics. For several plots, the x-axis is dimensionless and features an enumeration of the complexes. This enumeration is such that (i)the five groups are listed from left to right (ii)within a group, the complexes are ranked according to the interface size SASL. The plots concerned are: 11, 12, 13, 15(a,b), 17(a,b), 18. Zooming along the x-axis allows one to see the names of the complexes.



a.

Figure 10: Linear correlations between number of interface atom values of the AB (empty box) and ABW (full box) models versus Chakrabarti-Janin-2002 [CJ02]

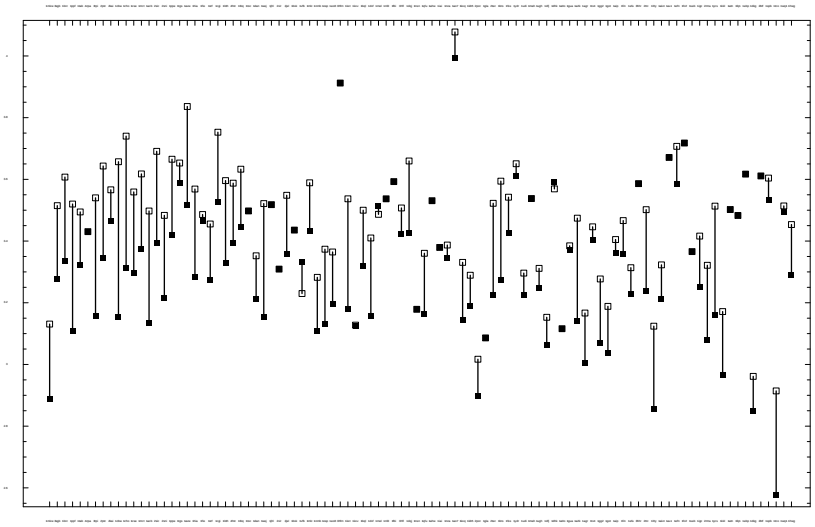


Figure 11: Average number of neighbours n_g in the AB model (empty box) and the ABW model (full box).

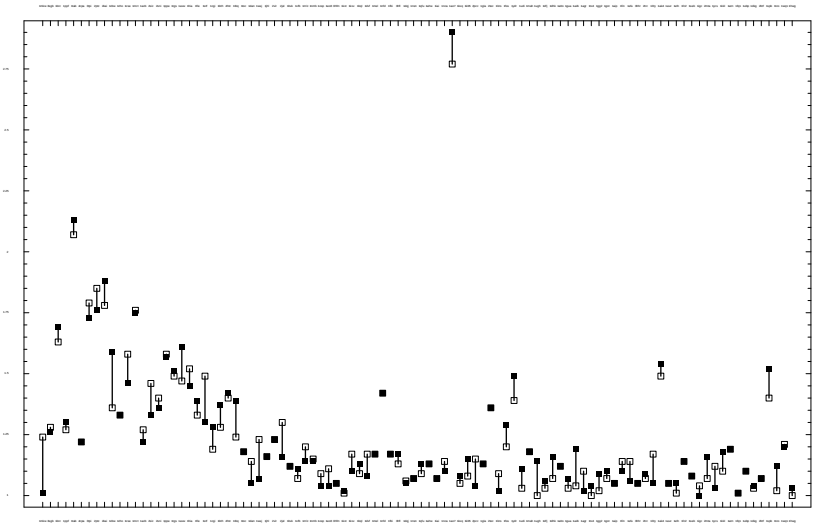


Figure 12: Comparison of r_{Mm} in the AB model (empty box) and the ABW model (full box).

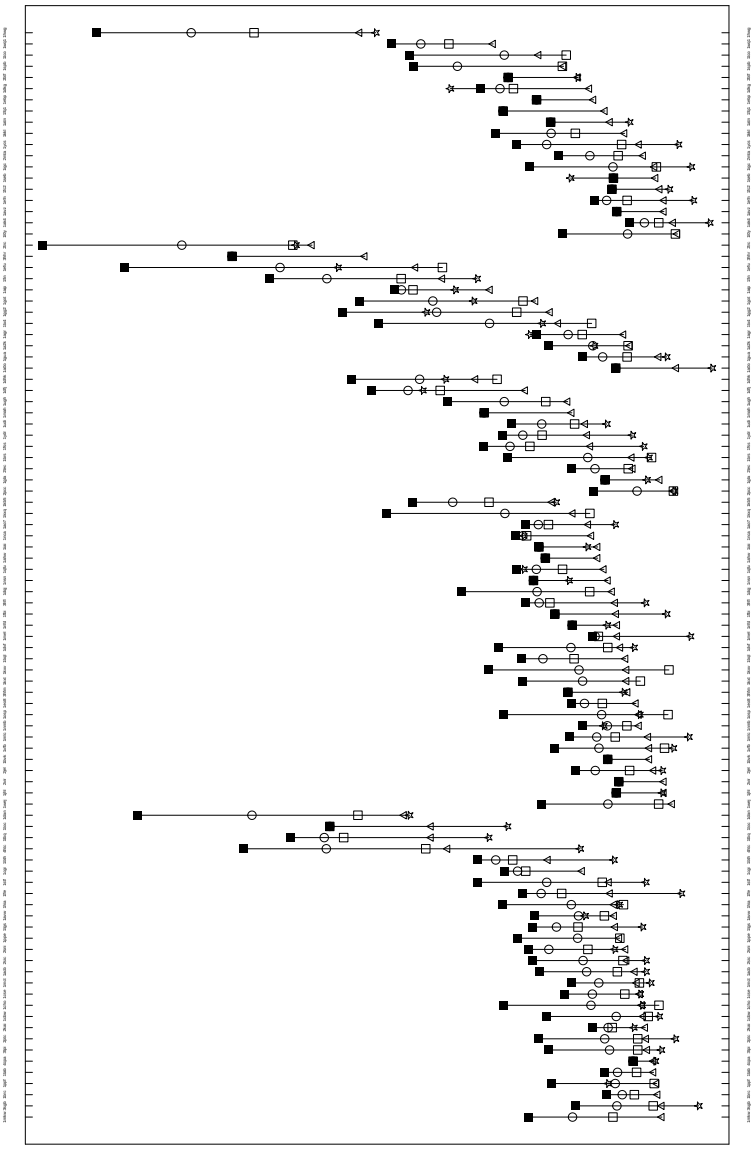
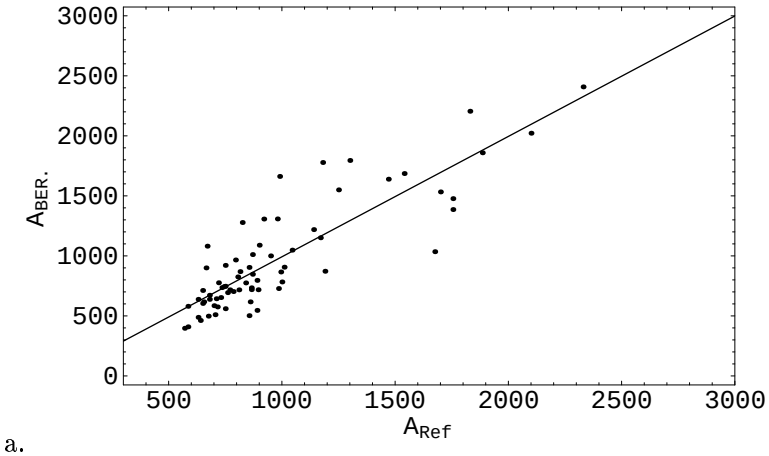
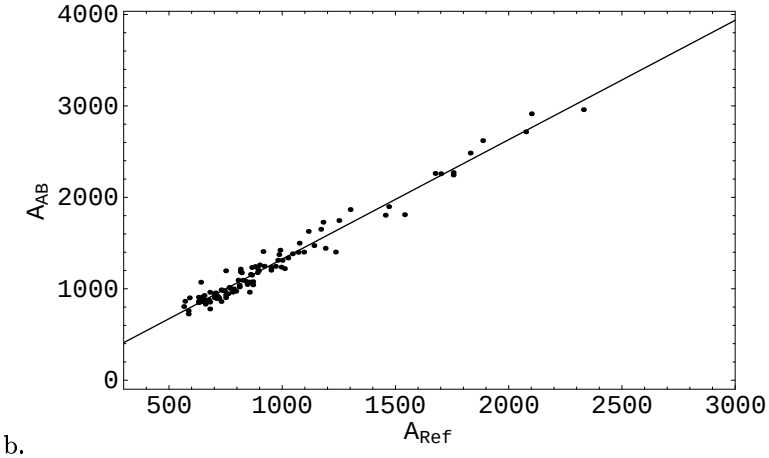


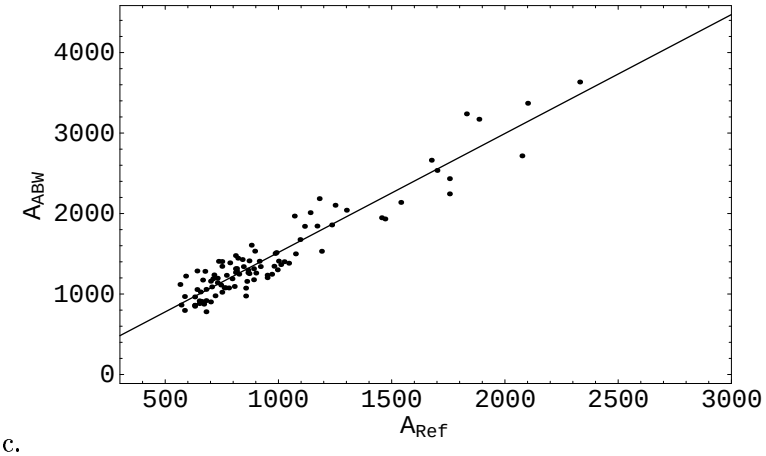
Figure 13: Comparison of the total surface area of bicolor interfaces. Empty circle (○): area A_{AB} in the AB model; Empty box (□): area A_{AB} in the ABW model; Full box, the area A_{ABW} in the ABW model. Star (⋆): values from Janin *et al.* [CJ99]; Triangles (△): values from by Ban *et al.* [BER04].



a.



b.



c.

Figure 14: Linear correlations between area values : (a)A_{BER} VISA vs $A_{Ref.}$ SASL (b)A_{AB} vs $A_{Ref.}$ SASL (c)A_{ABW} vs $A_{Ref.}$ SASL

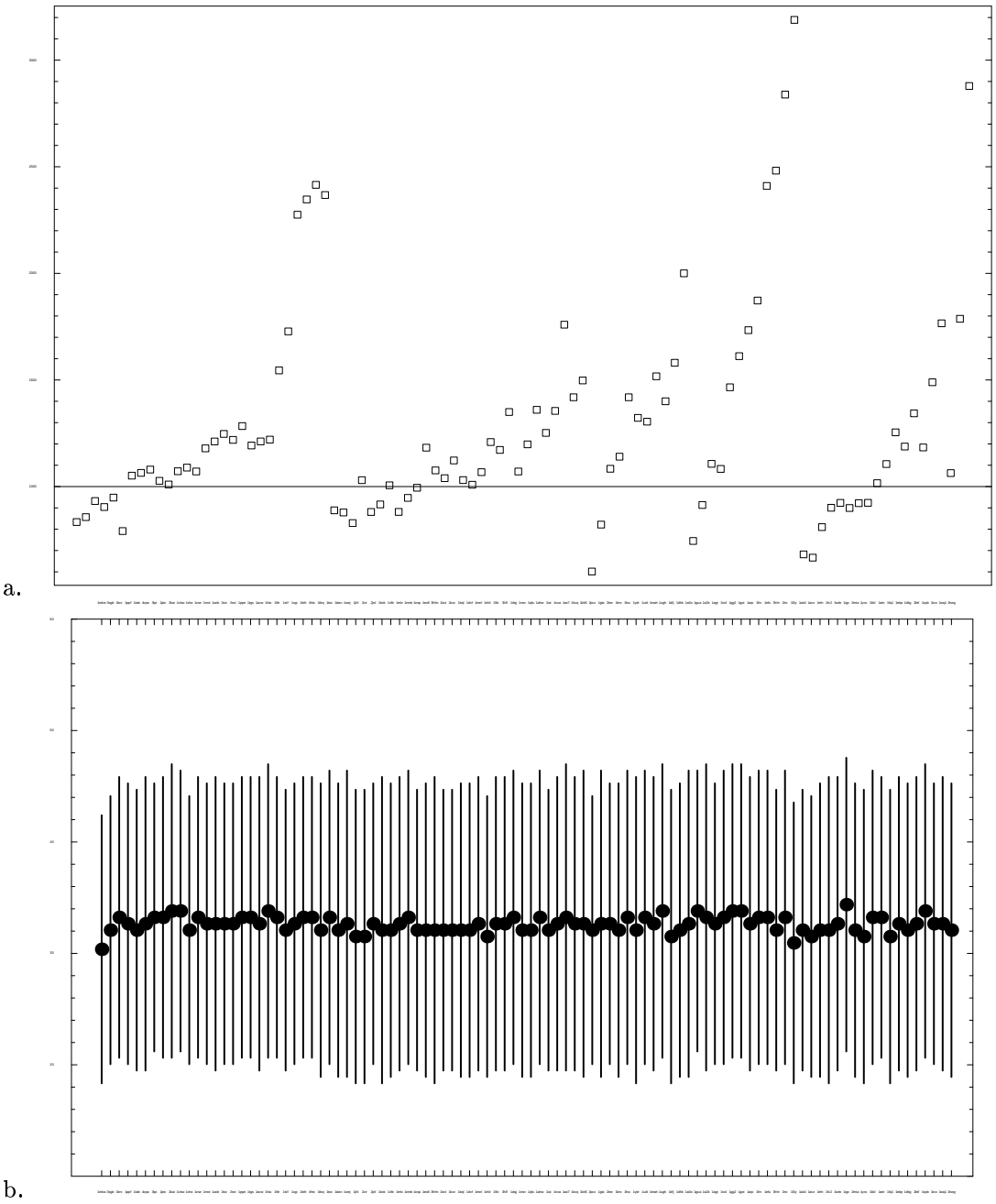


Figure 15: Curvature properties (a)Global measure i.e. s_H (b)Expectation and std deviation of the dihedral angle

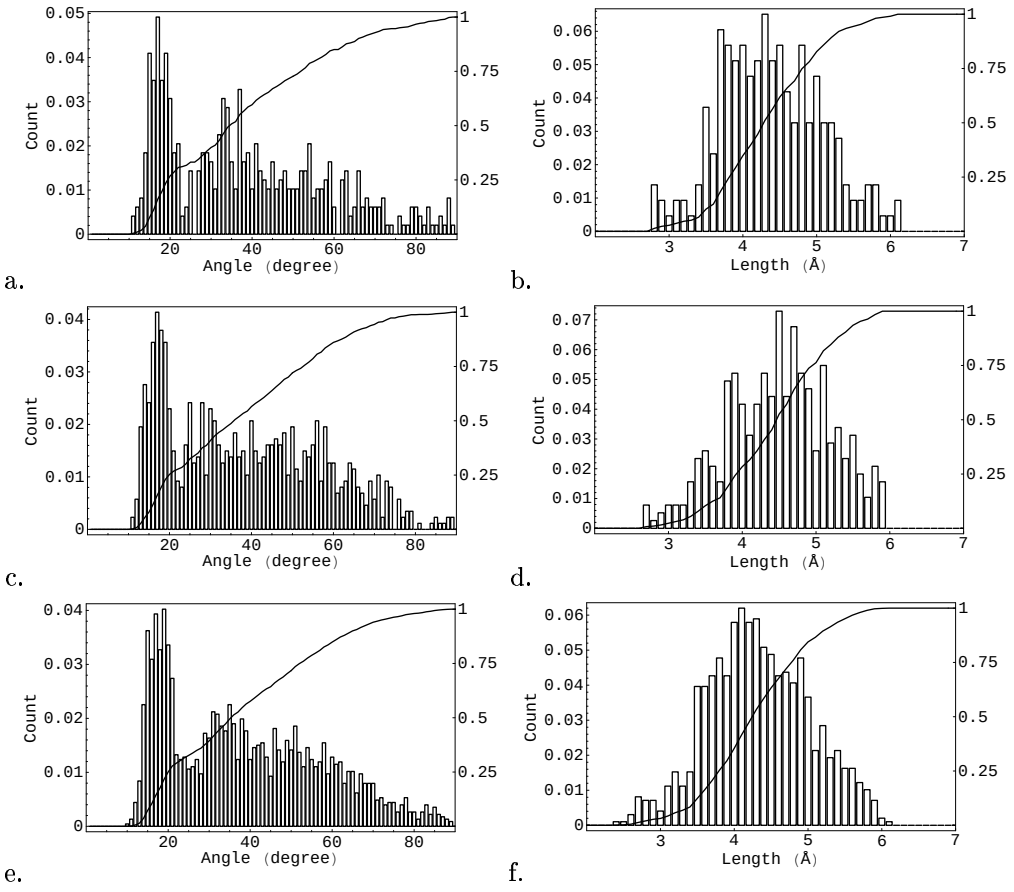
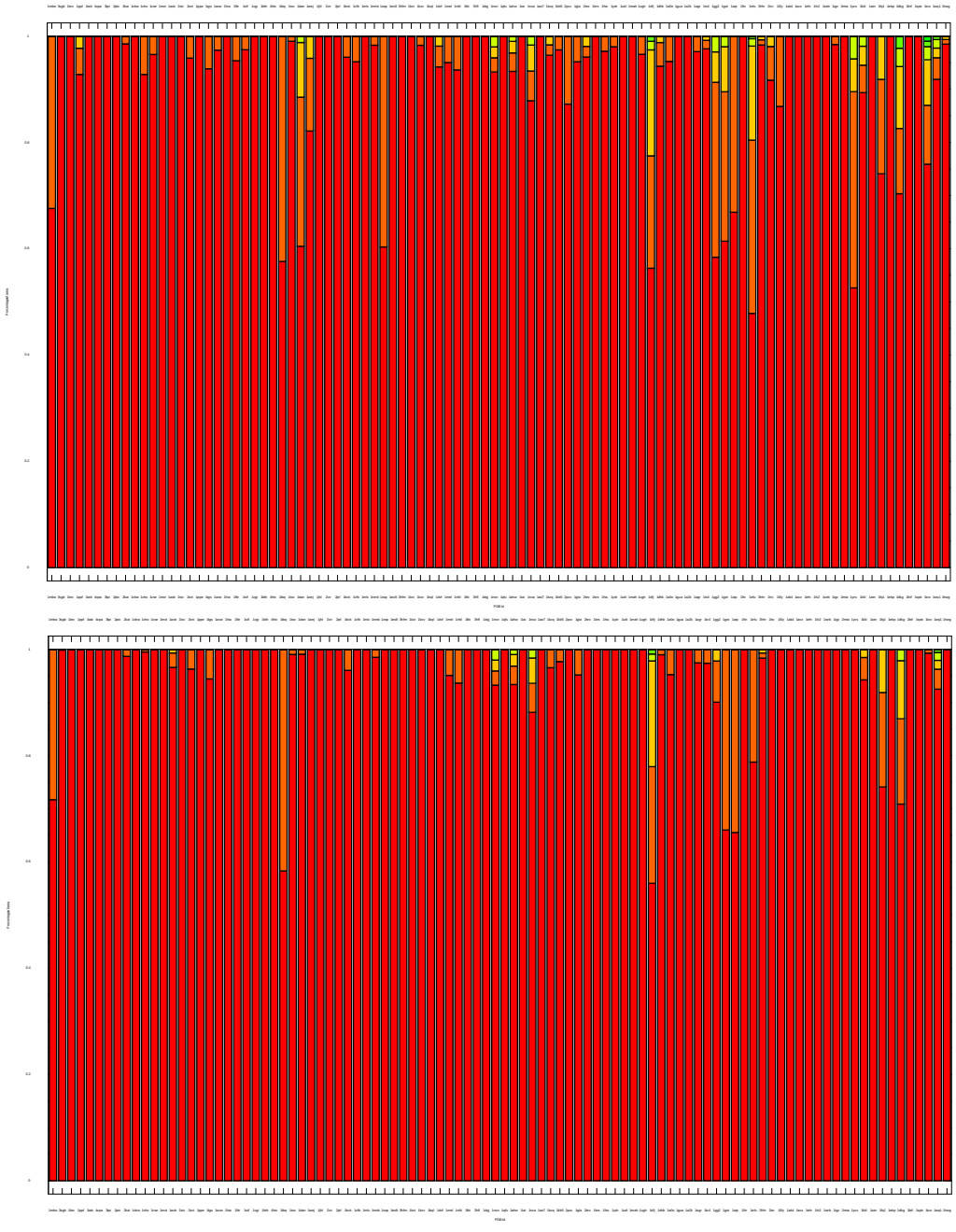


Figure 16: Distribution of angles and interface edge lengths for three selected complexes in the AB model: (a,b)1a0o (c,d)1udi (e,f)2trc.



a.

b.

Figure 17: Contribution of the number of connected components to the VISA area values A_{AB} and A_{ABW}: (a)AB model (b)ABW model after the merge step.

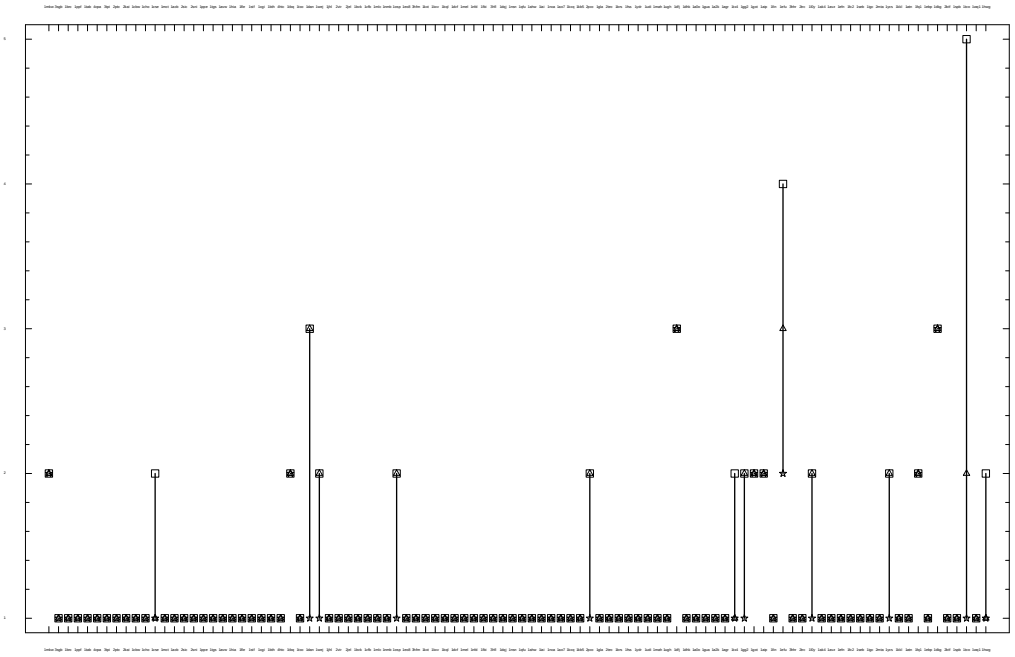


Figure 18: Number of significant connected components. Triangles: AB model: above 10% of the VISA of the AB interface. Squares: ABW model: above 10% of the VISA of the AB interface before the merge process. Stars: ABW model: above 10% of the VISA of the ABW interface after the merge process.

10 `intervor`: stand-alone program, VMD Plugin, Web site

Stand-alone program. `intervor`is a C++ program which can be called from the command line. Its arguments are the pdb file, the tags of the protein chains to be labeled as A and B, a boolean telling whether structural water is to be included, the radius of the water probe (default is 1.4Å), and the threshold M to be used to discard large facets —by default no facet is discarded.

VMD plugin. `intervor` can also be called from a VMD plugin, in which case the following objects are passed to VMD:

- the interface atoms. Atoms of type A (B)(W) are colored in light/deep blue (pink/red) (light/dark grey) depending on whether they are accessible or buried in the complex. These atoms are represented in two venues, namely with their Van der Waals radii, or expanded by the water probe.
- the bicolor interface AB and the interface $AW - BW$. Facets of the bicolor interface follow the color conventions listed in table 2. Facets of the $AW - BW$ interface are displayed in purple.
- the boundary loops of each interface (AB, $AW - BW$) before the merging step, and the boundary nets of the ABW interface. Different loops and nets are identified by different colors, but these color do not have any (chemical) meaning.

To ease the visual inspection of interfaces, all the objects mentioned above are independent VMD object and can be selected independently from the VMD main.

Web site. For the complexes of [CJ99, CMJW03, BCRJ04], interfaces can be retrieved from `www-sop.inria.fr/geometrica/team/Frederic.Cazals/intervor/index.html`. For each complex in each model (AB or ABW), we provide a VRML file for quick inspection, and a full VMD file for fine inspection.

Software setup. This experimental study has been conducted as follows. A suite of Perl script (1)call *intervor* on the complexes (2)generate the raw results files, in txt and LaTeX formats. The raw results are processed by Mathematica functions for the graphics and the statistics. Finally, the vmd interface files are generated by a Python script.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399